CAMPUS
DE EXCELENCIA
INTERNACIONAL

Graduado en Ingeniería Informática

Universidad Politécnica de Madrid

Facultad de Informática

TRABAJO FIN DE GRADO

Análisis de textos explicativos del significado de datos cuantitativos

Autor: Silvia Carolina Núñez Núñez

Director: Martín Molina González

MADRID, 09 JUNIO DE 2013

ÍNDICE

Resumen

En este Trabajo de Fin de Grado se ha realizado el análisis de textos explicativos de datos cuantitativos, con la finalidad de dar a conocer cuáles son las relaciones, basándose en la Teoría de la Estructura Retórica, entre las distintas frases de un texto de más común uso en documentos periodísticos relacionados con el comportamiento humano y el uso que hacen las personas de las redes sociales. Además de ello se han analizado un conjunto de 20 textos (alrededor de 1200 páginas) obteniendo frases típicas relacionadas con el mismo tema, que sirvieron como base para la construcción del modelo compuesto por un total de 101 patrones.

En un futuro, este Trabajo puede ser continuado, si así se desea, para lo cual se plantean las siguientes posibilidades:

- Ampliar el conjunto de patrones proporcionado.
- Construir un Sistema Generador de Textos automáticos basados en los patrones creados.
- Ampliar el estudio y extrapolarlo a diversos temas.

Keywords – Inteligencia Artificial, redes sociales, comportamiento humano, frases, patrón, texto, Lenguaje Natural, análisis, data journalism, bases de datos, sistemas de datos a textos.

Abstract

In this Final Project has been performed an analysis of quantitative data explanatory texts, in order to make known what are the relationships, based on Rhetorical Structure Theory, between the different sentences of a text of most common use in journalistic texts related to human behavior and the use people make of social networking. Furthermore have been analyzed a set of 20 texts (about 1200 pages) obtaining typical sentences related to the same topic that served as the basis for construction of the model consists of a total of 101 patterns.

In the future, this work can be continued, if so desired, for which the following possibilities are raised:

- Extend the set of patterns provided.
- Build an Automatic Text Generator System based on the patterns collected in this study.
- Expand the study and extrapolate it to various topics.

Keywords – Artificial Intelligence, social network, human behavior, templates, pattern, text, Natural Language, analysis, data journalism, datasets, data to text systems.

Introducción

Una de las metas más ambiciosas que se ha propuesto el mundo de la Ingeniería Informática, específicamente la rama de la Inteligencia Artificial, es crear máquinas inteligentes que no sólo sean capaces de reproducir texto, sino que también generen Lenguaje Natural e imiten el razonamiento del ser humano.

A lo largo de los años, se han obtenido muchos avances, como lo son la creación de *bots* conversacionales, que son máquinas calificadas para mantener una conversación haciendo uso de su base de conocimientos para dar una respuesta, o Sistemas Generadores de Lenguaje Natural, que parten prácticamente de la misma idea, es decir, hacer uso de bases de datos cuyo contenido suelen ser datos numéricos y en conjunto con su base de conocimientos generar frases.

A partir de este tipo de sistemas, surge lo que se conoce como *Data journalism,* un término utilizado para referirse al periodismo originado a partir de bases de datos y el uso de técnicas o sistemas informáticos que generen las noticias.

Basándose en la idea de este nuevo tipo de periodismo, el presente Trabajo de Fin de Grado, es un proyecto de investigación que se centra en el análisis de textos explicativos de datos cuantitativos, haciendo uso de bases de datos y artículos periodísticos que están relacionados con el tema elegido "El comportamiento humano y el uso que hacen las personas de las redes sociales".

El proceso de análisis está dividido en dos fases, la primera corresponde al estudio de textos periodísticos, apoyándose en la Teoría de la Estructura Retórica la cual ayuda a comprender las relaciones que existen entre las distintas partes del texto, de modo que se establecen conclusiones sobre cuáles relaciones tienen más presencia en los artículos vinculados con el tema elegido. La segunda parte está conformada por la selección de frases, siguiendo una metodología que ha sido definida a lo largo de la realización del Proyecto, que sirven de ejemplo en la creación de patrones que puedan ser utilizados en un futuro por un Sistema Generador de textos automáticos, o como base para la creación de un conjunto más amplio de patrones del mismo tema y/o sea extrapolado a otros tópicos.

Para el desarrollo del Trabajo de Fin de Grado se han propuesto un total de cinco objetivos:

1. Estudio del problema
2. Revisión de métodos
3. Búsqueda de fuentes documentales y textos periodísticos
4. Proceso de análisis

5. Evaluación del modelo

Los dos primeros objetivos hacen referencia a los conceptos que se consideran necesarios clarificar para la mejor comprensión del Proyecto y a trabajos han sido realizados previamente para conocer el avance o estado actual de los logros de la Inteligencia Artificial y la generación del Lenguaje Natural.

Los siguientes dos se basan completamente en el proceso de análisis, en cada una de las dos fases que fueron descritas previamente, es decir, en el análisis de textos para comprender las relaciones que más presentes están en el tipo de texto elegido y la construcción del conjunto de patrones.

Y por último, está la evaluación del modelo, que consiste en el establecimiento de métricas que resultan de interés, para hacer comparaciones y/o conclusiones del análisis realizado.

Capítulo 1

ESTADO DEL ARTE

1. ESTADO DEL ARTE

En esta sección se describen conceptos de gran importancia para la comprensión del Trabajo que se ha realizado, los cuales están relacionados con la Inteligencia Artificial y el periodismo.

Dentro del campo de la Inteligencia Artificial, se aclara el significado y relevancia del Procesamiento del Lenguaje Natural. Además, se describen proyectos previos desarrollados de *Data to Text Systems* y los *bots* conversacionales que representan los primeros avances obtenidos.

Por otro lado, se explica el término que es utilizado hoy en día para referirse al periodismo de la nueva era, *data journalism* y cómo movimientos o iniciativas como *Open data* están relacionados.

1.1. Conceptos básicos

Para entender mejor el objetivo del Trabajo, es bueno conocer los fundamentos en los cuales está basado, para ello es importante definir dos grandes campos los cuales son: Procesamiento del Lenguaje Natural (PLN) y *Data Journalism.* De igual manera se define el movimiento *Open data*, el cual es clave en el desarrollo de este Trabajo.

1.1.1. Procesamiento del Lenguaje Natural (PLN)

La más valiosa posesión que tiene el ser humano es el conocimiento, el cual puede ser obtenido mediante distintas fuentes, como lo son: periódicos, libros, informes, medios de comunicación, e incluso la comunicación entre personas, entre otros. Pero realmente lo que hace este conocimiento tan apreciable, es la capacidad que tiene el individuo de buscar la información, compararla con distintas fuentes, sacar sus propias conclusiones y manejar la información como lo desee.

Si bien es cierto que gracias a los avances tecnológicos, los ordenadores hoy en día son capaces de procesar más información de lo que puede una persona en toda su vida. Sin embargo, el reto más grande al que se enfrenta la informática es el de poder construir máquinas que sean capaces de seleccionar cierta información de su Base de Conocimiento, organizarla, y determinar cómo producir el texto en Lenguaje Natural.

El Lenguaje Natural puede definirse como el lenguaje utilizado por los seres humanos bien sea hablado o escrito, con el propósito de comunicarse. Está conformado por una sintaxis, que es el conjunto de reglas que garantizan que está siendo correctamente utilizado o construido gramaticalmente y la semántica la cual garantiza que tenga sentido.

Hasta el momento los ordenadores sólo son capaces de procesar información, pero no entenderla, es decir, para la máquina sólo representa una cadena de caracteres que carece de sentido.

Por éste motivo surge lo que hoy se conoce como Procesamiento del Lenguaje Natural [Carbonell, 1992], en inglés *Natural Language Processing*, que es una disciplina de la Inteligencia Artificial y una rama de la ingeniería que se basa en la creación de mecanismos que permitan establecer la comunicación entre persona-máquina.

Para poder hablar de Procesamiento de Lenguaje Natural, también hay que hablar de Generación del Lenguaje Natural, que es un término utilizado para hacer referencia al proceso de creación de frases en Lenguaje Natural, bien sean habladas o escritas, para poder establecer la comunicación.

1.1.2. Data Journalism

Data Journalism o Data Driver Journalism [Bradshaw, 2013], es un término utilizado para hacer referencia al periodismo de la nueva era, es decir, aquel que proviene de la combinación del uso de datos numéricos y la informática para generar noticias.

En la figura 1, se puede observar el proceso *Data Journalism* para la generación de noticias.

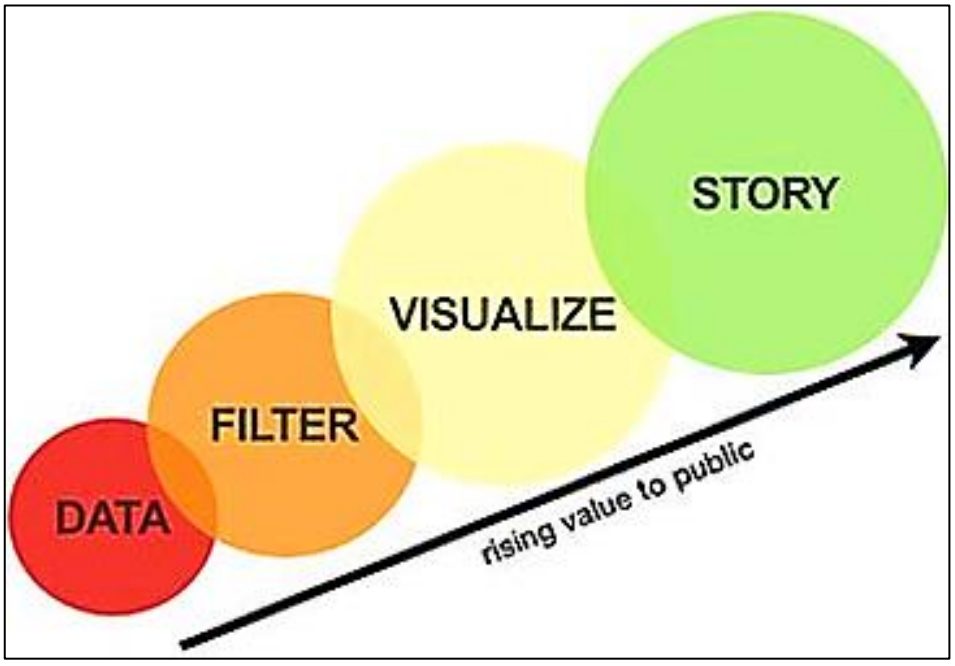


Figura 1: *Proceso de Data Journalism.*

El primer paso de este proceso es encontrar los datos que están disponibles libremente para su uso en la web, que gracias a iniciativas como *Open Data* es posible. Luego es necesario filtrar dicha información, es decir, elegir bien los datos que quieren ser utilizados para luego finalmente visualizar que es lo que se quiere transmitir y generar la noticia. En la actualidad éste proceso se realiza manualmente, todavía no hay un sistema que automatice la generación de noticias.

1.1.3. Open data

Desde hace unos años atrás han surgido movimientos como *Open source* [OSI, 1999], *Open hardware* [Suehle, 2011], *Open content* [OPL, 1998] y *Open access* [Harnad, 2005], con una finalidad en común y es que toda información o herramienta esté al alcance de todas las personas.

Open data [OKF, 2012] es un iniciativa que nace con la misma idea que los movimientos anteriormente nombrados y gracias a la gran popularidad que ha tenido Internet y *World Wide Web* hoy en día está teniendo mucho auge.

A este éxito se puede añadir también el desarrollo de proyectos como: *data.gov* de Estados Unidos o *data.gov.uk* del Reino Unido, de los cuales muchos países se han contagiado, estos proyectos publican ciertos datos como: datos estadísticos, fórmulas matemáticas, físicas o químicas, datos médicos, entre otros, que pueden resultar interesantes para todo público.

La finalidad *Open data* es que las personas puedan acceder a estos datos y puedan reutilizarlos sin tener ningún inconveniente con el *copyright*, las *patents* o con cualquier otro mecanismo de control ante el plagio.

1.2. Sistemas Generadores de Lenguaje Natural

A continuación se describen sistemas que generan descripciones de datos numéricos, así como programas que hacen uso de bases de conocimientos para poder mantener una conversación persona-máquina.

1.2.1. Sistemas de Datos a Textos

Los Sistemas de Datos a Textos o mejor conocidos en inglés como *Data to text Systems* [Reiter, 2007], son sistemas que generan resúmenes textuales de datos procedentes de bases de datos o *datasets*. El propósito de estos sistemas es crear textos explicativos para que las personas puedan comprender mejor el significado de los datos numéricos.

Un ejemplo de texto generado, se puede observar en la figura 2, en la cual el sistema recibe como datos de entrada los 6 niveles de polen en diferentes partes de Escocia, y a partir de ellos genera el siguiente fragmento.

```
Grass pollen levels for Tuesday have decreased from the high levels
of yesterday with values of around 4 to 5 across most parts of the
country. However, in South Eastern areas, pollen levels will be high
with values of 6.
```

Figura 2: *Ejemplo de Data to Text Systems.*

Uno de los beneficios obtenidos al utilizar este tipo de sistemas es el ahorro de tiempo, ya que no es necesario que las personas dediquen muchas horas de trabajo en la interpretación de datos para poder generar un texto, por el contrario, pueden hacer uso del sistema y verificar si el texto generado es correcto.

Para tener un conocimiento más amplio de los *Data to Text System*s, se nombran y explican brevemente a continuación algunos de los sistemas que se utilizan para la interpretación de datos numéricos en distintas áreas.

1.2.1.1. BabyTalk Project

BabyTalk Project [Moncur, 2008] es una herramienta desarrollada por el Departamento de Ciencias de la Computación de la Universidad de Anderbeen, la Escuela de Salud en Ciencias Sociales y el Departamento de Vida Infantil y Neurociencia Cognitiva de la Salud Humana de la Universidad de Edimburgo.

BabyTalk es utilizada por la Unidad de Cuidados Intensivos de Neonatales (siglas en inglés, NICU) de Edimburgo, para generar informes a partir de datos medidos, con la finalidad de dar más apoyo a la decisión de los médicos basada en el tratamiento y estudio de la información recogida y que los padres estén más al corriente de la situación de su(s) hijo(s).

El modo de funcionamiento de esta herramienta se basa en la recogida e interpretación de los datos relacionados con los eventos fisiológicos de cada bebé, a partir de esos datos se procede a la creación de resúmenes y a la personalización de los mismos teniendo en cuenta a quien serán dirigidos, ya que los médicos, enfermeras, y padres de los pacientes no poseen los mismos conocimientos técnicos.

Esta herramienta es muy completa ya que se apoya en otros tres sistemas que hacen que toda la información generada sea fiable y de fácil acceso a todas las partes interesadas. Dichos sistemas son:

- ***BT-Doc:*** genera los resúmenes médicos durante largos periodos de tiempo.
- ***BT-Nurse:*** genera resúmenes médicos durante cortos periodos de tiempo que pueden resultar interesantes para el cambio de guardia de las enfermeras.

- ***BT-Family:*** genera resúmenes médicos en un lenguaje menos formal o técnico para que sea de fácil comprensión para los familiares del paciente.

1.2.1.2. ANA

ANA [Nevin/Johnson, 2002], es una herramienta pionera generadora de textos, desarrollada por Karen Kukich en el año 1983, que resume en cortas oraciones las fluctuaciones de distintos índices del mercado en periodos de tiempo de cada media hora.

ANA es especialmente conocida porque los resúmenes o reportes que genera son de muy buena calidad, ya que los compara con los realizados por los seres humanos. Sin embargo, tiene una limitación: no transmite información histórica, es decir, produce resúmenes de la actualidad pero no es capaz de relacionarla con los realizados en semanas o meses anteriores.

Dicha herramienta está formada por cuatro grandes módulos que trabajan en conjunto para poder sintetizar información a partir de datos numéricos, los cuales son:

1. Generador de hechos.
2. Generador de mensajes.
3. Organizador de discursos.
4. Generador de textos lingüísticos.

El primer módulo genera hechos abstractos a partir de datos numéricos, que luego son utilizados para generar cortas frases que pueden resultar de interés para el público o audiencia. Posteriormente, estos mensajes se juntan en un mismo párrafo con sentido y bien organizados para formar textos acerca un tema determinado.

En la figura 3, se puede observar un fragmento de texto generado por ANA.

The stock market was catapulted sharply and broadly higher yesterday, as stock prices posted gains for most of the day. Trading was active.

Figura 3: *Fragmento de texto generado por ANA.*

1.2.1.3. SumTime

SumTime [EPSRC, 2005], es un proyecto desarrollado con la finalidad de implementar técnicas que generan resúmenes de datos numéricos sobre meteorología. Éste proyecto es el que tiene más similitud con el Trabajo desarrollado debido a la metodología seguida.

Dicha metodología comprende la búsqueda y estudio de textos relacionados con el tema para identificar frases de más común uso, y la recolección de datos numéricos de meteorología. Las frases comunes y los datos numéricos luego son insertados en una base de conocimientos para ser utilizados en la generación automática de textos a partir de datos numéricos.

En la figura 4, se puede observar dos tablas en las que se almacenan palabras y datos numéricos para luego generar el texto descriptivo.

Words		
id	word	pos
1080713329	E	0
1080713329	SE	1
1080713329	18	2
1080713329	22	3
1080713329	GRADUALLY	4
1080713329	BACKING	5
1080713329	EASING	6

MMOData							
filename	Clow	Temp	Wind	Precip	Snow	Thunder	Lapse
10Apr2001_09.mmo	8	8	9	0,182	-21,5	35,5	7,64
10Apr2001_09.mmo	8	8,1	10	0,197	-21,8	35,8	7,66
10Apr2001_09.mmo	8	8,1	10	0,211	-22,1	36	7,69
10Apr2001_09.mmo	7,97	8,2	11	0,222	-22,4	36,4	7,72
10Apr2001_09.mmo	7,85	8,3	11	0,229	-22,7	36,8	7,74
10Apr2001_09.mmo	7,81	8,3	12	0,231	-23	37,5	7,77
10Apr2001_09.mmo	7,56	8,5	12	0,195	-21,5	31,6	7,6

Figura 4: *Tablas de palabras y datos numéricos.*

1.2.2. Bots conversacionales

Uno de los primeros logros o avances de la informática en el mundo de la Inteligencia Artificial, son los conocidos *bots* conversacionales, que hacen uso de Bases de Conocimientos, para poder entablar una conversación persona-maquina. Aunque para el desarrollo de éste trabajo, no es muy importante saber cómo funcionan, nunca está de más conocer del tema.

1.2.2.1. Eliza

Eliza [Busacca, 1998] es un programa informático realizado durante los años 1994-1996 por Joseph Weizenbaum. Es uno de los primeros programas en procesar Lenguaje Natural. Dicha herramienta puede mantener conversaciones coherentes, por *chat*, con un individuo cualquiera.

En la figura 5, se puede observar el funcionamiento.

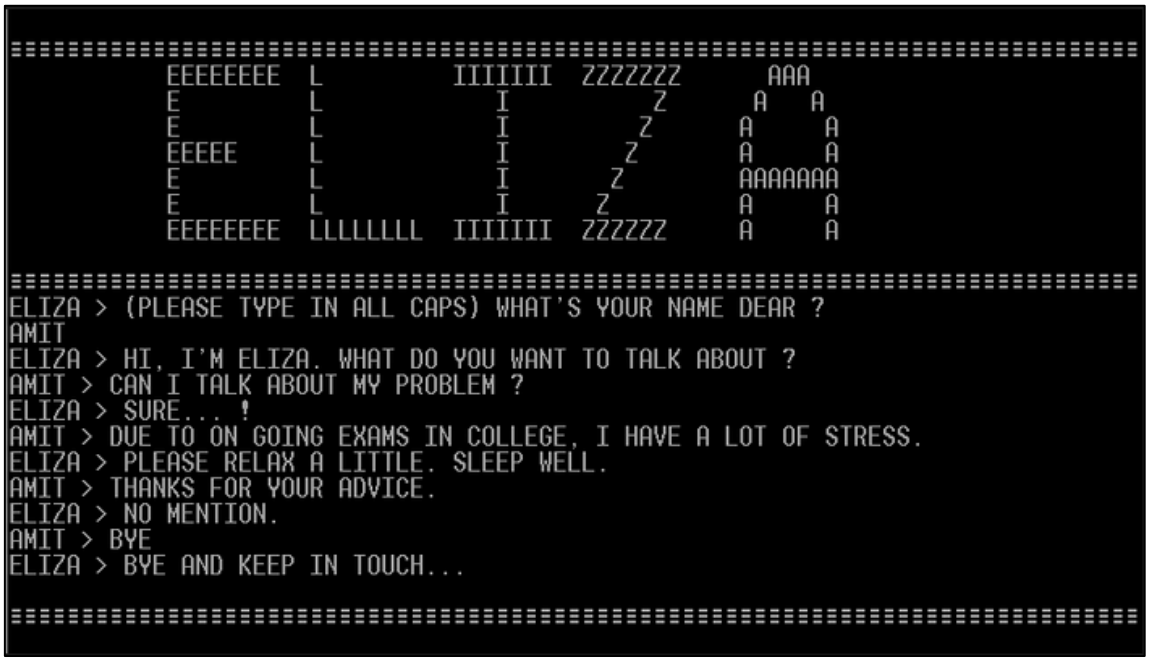


Figura 5: *Conversación con Eliza.*

El modo de funcionamiento es simple, cuando el programa recibe una línea de texto, busca palabras claves dentro de la frase y responde con frases modelo que están en su Base de Conocimientos.

Pero esta herramienta es limitada, si recibe alguna frase que no es capaz de reconocer, coge esa misma frase y la reordena a modo de pregunta, lo que hace que la conversación se vuelva poco coherente o se pierda el hilo de la misma.

1.2.2.2. ALICE

ALICE [Wallace, 2013] acrónimo de *Artificial Linguistic Internet Computer Entity*, es un proyecto desarrollado por Internet que se basa en la creación de *bots* conversacionales, muy similar a Eliza, que lo que hace es mantener conversaciones inteligentes con diversos usuarios a través de un *chat*, es tan amplia su Base de Conocimientos, que muchas veces es muy difícil darse cuenta que se está hablando con un robot.

En la figura 6, se puede observar el funcionamiento.

Figura 6: *Conversación con ALICE.*

Sin embargo también tiene limitaciones, cuando una palabra está mal escrita o una frase no tiene coherencia, no es capaz de entenderla y al igual que Eliza utiliza la frase reordenada para hacer preguntas o reflexionar acerca de lo dicho.

1.2.2.3. Cleverbot

Cleverbot [Saenz, 2010], es un programa de Inteligencia Artificial, mucho más avanzado que Eliza y ALICE, ya que es capaz de mantener conversaciones con personas en diferentes idiomas, además de aprender de las charlas previas que ha tenido con anteriores usuarios, por lo cual su Base de Conocimientos es más amplia que cualquier otro *bot* conversacional.

En la actualidad, la aplicación está disponible para dispositivos móviles como IPhone y Android, la cual ha tenido mucha fama entre los usuarios, porque no sólo se limita al chat mediante intercambio de frases escritas, sino que también se puede entablar una conversación haciendo uso de la propia voz.

La aplicación no es gratuita, cuesta 0.99$ en EE.UU., 0.79€ en Europa y 0.69£ en el Reino Unido, lo cual no ha implicado ser un obstáculo para que las personas hagan uso de ellas.

1.2.3. Resumen

Una vez conocidos algunos *Data To Text Systems* y obtenido un poco de cultura con los primeros avances de la Inteligencia Artificial con los *Bots* Conversacionales, queda claro cuál es el objetivo principal de la informática en los últimos tiempos y es que las

máquinas sean capaces de no sólo generar Lenguaje Natural, sino aprender de él, para por si solos obtener un conocimiento al igual que un ser humano.

De los *Data to Text Systems*, el que más se asemeja en cuanto a metodología de trabajo es *SumTime*, porque como ya se ha comentado anteriormente el objetivo de este Proyecto es realizar un análisis de textos explicativos de datos cuantitativos, y una de las fases de trabajo de *SumTime* es justamente esa, recopilar información numérica y textual de un tema específico para luego generar un modelo que pueda ser utilizado para generar textos explicativos.

Lo que diferencia al análisis que se ha realizado, es que se fundamenta en la Teoría de la Estructura Retórica para conocer las relaciones que suelen estar presentes en los textos periodísticos seleccionados, y la construcción de un modelo formado por patrones que puedan ser utilizados en un futuro por *data to text system*s.

1.3. Teoría de la Estructura Retórica

Un grupo de investigadores del *Information Sciences Institute*, de la Universidad del Sur de California, mientras trabajaban en la auditoria de textos mediante un computador, se dieron cuenta que no existía ninguna teoría que estudiara detalladamente la estructura de los textos para la generación automática de los mismos.

Por esta razón, hacia el año 1983, se planteó la Teoría de la Estructura Retórica [Mann/Taboada, 2012], en inglés *Rhetorical Structure Theory*. La cual fue desarrollada a partir del estudio y análisis de la estructura de un conjunto de textos provenientes de distintas fuentes. Durante el estudio, los investigadores notaron que las oraciones que conforman un párrafo o texto, no son un conjunto de ideas aisladas sino que tienen un "algo" que hace que se unan, ese "algo" es la coherencia, que puede definirse como la propiedad que tienen los textos que hace que se puedan entender como una única entidad.

1.3.1. Definiciones de Estructura, Esquemas y Relaciones

La Teoría de la Estructura Retórica, está basada en cuatro principales elementos que son definidos independientemente del contexto en el que sean utilizados [Mann/Thompson, 2006]. Dichos componentes son: estructura, esquema, aplicaciones de esquema y relaciones.

1.3.1.1. Relaciones

Como su nombre lo indica define las relaciones que puede haber entre dos partes de un texto que normalmente están adyacentes, estas son relaciones de núcleo-satélite.

En la tabla 1, se pueden observar las relaciones más comunes.

Nombre de la relación	Núcleo	Satélite
Circunstancia ***(Circumstance)***	texto que expresa los hechos o ideas que tienen lugar en el marco contextual	marco contextual, situacional o temporal
Solución ***(Solutionhood)***	una situación o método que satisface la necesidad, total o parcialmente	una pregunta, petición, problema u otra necesidad
Elaboración ***(Elaboration)***	información básica	información adicional
Fondo ***(Background)***	texto del cual se facilita la comprensión	texto que facilita la comprensión
Capacitación ***(Enablement)***	una acción	información que ayuda al lector a llevar a cabo la acción
Motivación ***(Motivation)***	una acción	información que pretende conseguir que el lector desee llevar a cabo la acción
Evidencia ***(Evidence)***	una afirmación	información que pretende conseguir que el lector esté de acuerdo con la afirmación
Justificación ***(Justify)***	Texto	información que apoya el derecho del escritor a escribir el texto
Causa Voluntaria ***(Volitional Cause)***	una situación	otra situación que causa la primera, mediante la acción voluntaria de una persona o personas
Causa Involuntaria ***(Non-volitional Cause)***	una situación	otra situación que produce la primera, sin ser causada por una acción voluntaria
Resultado Voluntario ***(Volitional Result)***	una situación	otra situación provocada por la primera, mediante la acción voluntaria de una persona o personas
Resultado Involuntario ***(Non-volitional Result)***	una situación	otra situación provocada por la primera, sin ser provocada por una acción voluntaria
Propósito ***(Purpose)***	una situación intencional	intención que se halla detrás de la situación
Antítesis ***(Antithesis)***	ideas que el autor apoya	ideas que el autor no apoya
Concesión ***(Concession)***	situación que el autor afirma	situación, de inconsistencia aparente con respecto al núcleo, pero que el autor también afirma
Condición ***(Condition)***	acción o situación que resulta de la situación condicionante	situación condicionante
Alternativa (anti condicional) ***(Otherwise)***	acción o situación que resulta de la no presencia de la situación condicionante	situación condicionante
Interpretación ***(Interpretation)***	una situación	interpretación de la situación
Evaluación ***(Evaluation)***	una situación	comentario que evalúa la situación
Reformulación ***(Restatement)***	una situación	reformulación de la situación

Nombre de la relación	Núcleo	Satélite
Resumen ***(Summary)***	Texto	un breve resumen del texto
Preparación ***(Preparation)***	texto que va a ser presentado	texto que prepara al lector para anticipar e interpretar el texto que va a ser presentado
Contraste ***(Contrast)***	una opción	la otra opción

Tabla 1: *Relaciones más comunes entre dos partes de un texto.*

Además de las relaciones mostradas en la tabla anterior que siguen la estructura núcleo-satélite, hay relaciones que tienen más de un núcleo y son denominadas relaciones multinucleares.

En la tabla 2, se pueden observar las relaciones multinucleares.

Nombre de la relación	Unidad	Otra unidad
Contraste ***(Contrast)***	una opción	la otra opción
Lista ***(List)***	un elemento	siguiente elemento
Secuencia ***(Sequence)***	un elemento	siguiente elemento
Unión ***(Joint)***	(sin requisitos)	(sin requisitos)

Tabla 2: *Relaciones multinucleares.*

A su vez, cada relación puede ser definida en 4 campos que son:

- Condiciones en el núcleo.
- Condiciones en el satélite.
- Condiciones en la combinación núcleo-satélite.
- El efecto.

En la tabla 3, se puede observar cómo estos campos aportan ciertos juicios que ayudan a la persona que analiza los textos en la construcción de la estructura.

Condición	
Elemento de definición	Conclusiones del analista
condiciones en el núcleo, N:	(ninguna)
condiciones en el satélite, S:	S presenta una situación hipotética, futura, o aún no realizada (con relación al marco contextual de S)
condiciones en la combinación de N + S	La realización de la situación presentada en N depende de la realización de la situación presentada en S
el efecto (la intención del autor al utilizar esta relación para dirigirse al lector; nunca está vacío)	El lector comprende cómo la realización de la situación presentada en N depende de la realización de la situación presentada en S

Tabla 3: *Condiciones de las relaciones.*

1.3.1.2. Esquemas

Los esquemas son estructuras que están definidas por las relaciones, que ayudan a comprender mejor como están estructurados los textos, es decir, como una frase puede estar conectada con el resto del texto. En la figura 7, se puede observar los cinco tipos de esquema en los cuales las líneas curvas representan las relaciones mantenidas entre una frase y otra dentro del texto y las líneas rectas el tramo nuclear del texto.

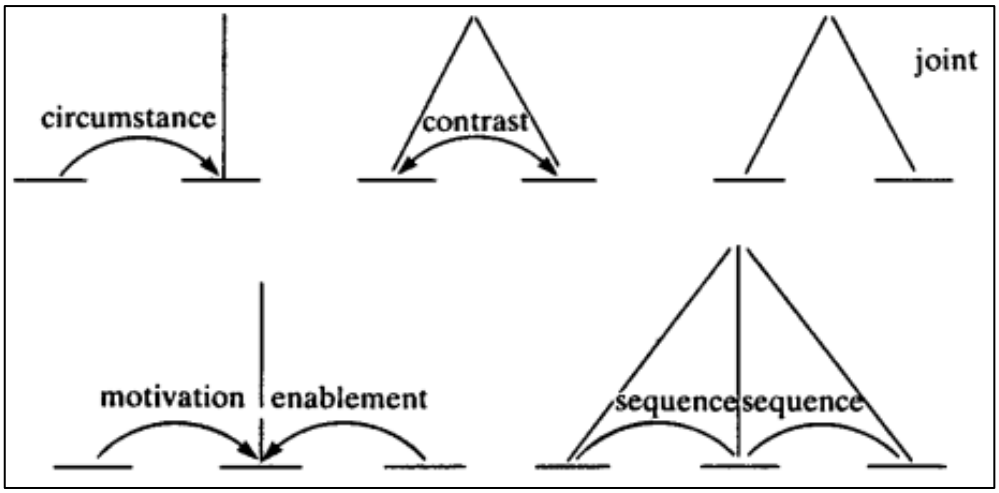


Figura 7: *Cinco tipos de esquemas.*

1.3.1.3. Aplicaciones de Esquemas

No siempre los esquemas que están definidos aparecen exactamente en el texto, sino variaciones de ellos, por este motivo se han llegado a las siguientes convenciones:

- **Tramos desordenados:** el esquema que es aplicado no limita en qué orden debe aparecer el núcleo y satélite en el tramo del texto.

- **Relaciones opcionales:** cuando los esquemas tienen múltiples relaciones, no está definido el número de relaciones que lo conforman ni de qué tipo son, sólo que debe tener al menos una.
- **Relaciones repetidas:** si la relación pertenece al esquema que es aplicado, puede ser utilizada el número de veces que sea necesario.

1.3.1.4. Estructuras

Para poder estudiar con más facilidad los textos, la mejor solución es dividirlos en unidades, pero no se trata de seleccionar un texto y separarlo en 3 o 4 partes, sino de ver cuales unidades son independientes de otras y siguen teniendo un sentido.

Una estructura puede ser definida como un conjunto de aplicaciones de esquemas que tienen ciertas limitaciones, que son:

1. **Completitud:** pequeños tramos de textos que constituyen una unidad entera.
2. **Conectividad:** aunque es un pequeño tramo de texto debe estar conectado para poder formar una unidad que pueda ser analizada.
3. **Unicidad:** se refiere a que una única relación puede ser aplicada al tramo del texto seleccionado.
4. **Proximidad:** el conjunto de tramos seleccionados puedan formar un texto con sentido.

Sin embargo, el conjunto unido de éstas sirven para hacer el análisis de textos con la teoría de la estructura retórica mediante árboles de decisión.

1.3.2. Relación con la Lingüística Computacional

En la actualidad, la Teoría de la Estructura Retórica es válida y utilizada en áreas como: el análisis de discursos, la lingüística teórica, la psicolingüística y la lingüística computacional.

En la Lingüística Computacional [Mann/Taboada, 2006], dicha teoría ha sido utilizada para la generación, procesamiento y síntesis de textos, la evaluación de argumentos, la traducción automática y por último pero no menos importante, siendo el uso más frecuente, la generación de Lenguaje Natural.

Sin embargo, para generar textos no sólo son necesarias las relaciones retóricas, también hay que tener conocimiento del dominio de la comunicación, es decir, saber cómo expresar los hechos e intenciones del tema determinado del cual se desee generar el texto. Dicho conocimiento puede ser obtenido en muchos casos por la lectura general del tema específico del cual se desee hablar.

Capítulo 2

PROCESO DE ANÁLISIS

2. PROCESO DE ANÁLISIS

En esta sección se plantea la metodología de trabajo definida y seguida durante la realización del Proyecto, tanto para la búsqueda y recolección de textos periodísticos y bases de datos como para el análisis de textos y construcción del modelo.

También se muestra un análisis de texto fundamentado en la Teoría de la Estructura Retórica, haciendo uso de una herramienta denominada *RSTTool*, la cual facilita la generación grafica del análisis. Por otro lado, se ha construido un modelo conformado por un total de 101 patrones, tomando como base oraciones obtenidas de los textos analizados.

Debido a las futuras líneas de trabajo que se han planteado para este Trabajo de Fin de Grado y que se explican más adelante, es importante dejar claro el público al que va dirigido dicho análisis. La idea es que este análisis se tome como ejemplo para futuros estudios o que los patrones construidos se utilicen como base de conocimientos de un sistema generador de textos o sea ampliado el conjunto de los mismos, por lo que se identifican dos públicos distintos:

1. Investigadores que deseen profundizar más en el tema.
2. Individuos que deseen utilizar los patrones como base de conocimientos (que no tienen por qué poseer conocimientos técnicos).

2.1. Fuentes documentales y bases de datos

Para la selección de fuentes documentales y bases de datos, que estén relacionados con el tema elegido, el comportamiento humano y el uso que hacen las personas de las redes sociales, es necesario tener bien definidos los objetivos que se desean alcanzar.

Lo más importante es conseguir que tanto textos como bases de datos estén relacionados directamente para que haya coherencia en el análisis realizado, es decir, hacer una búsqueda exhaustiva en diferentes sitios web, como pueden ser *data.gov, data.gov.uk*, o cualquier otra fuente en la que estén disponibles libremente los datos.

El motivo por el cual se hace tanto énfasis en esa relación, es que cuanto mejor se adapte el texto a los datos encontrados, mejor son los resultados de los patrones creados, lo cual agrega más valor al Trabajo realizado.

En concreto en este Trabajo se han seleccionado un total de 20 textos periodísticos (aproximadamente 1200 páginas) y 16 bases de datos, de los cuales tienen relación directa ocho textos y bases de datos.

A pesar de que toda la información recogida no está evidentemente vinculada, todos los textos seleccionados están redactados en base a datos numéricos, aunque para algunos de ellos no se pudo obtener los *datasets*. Sin embargo, se han elegido textos que contuviesen información parecida a los textos que si guardan relación directa con las bases de datos encontradas, para evitar grandes discrepancias.

De modo que en un futuro si se desea hacer comparaciones entre un texto generado automáticamente frente a uno escrito por un ser humano, se estima que el margen de error sea el menor posible.

2.2. Decisiones

Para los análisis de los textos se ha decidido no contemplar el conjunto completo de textos estudiados, debido a lo extensos que son, es por ello que se han estudiado 6 de los 20 artículos en total o las partes más representativas de los mismos, lo que simboliza el 30% del total.

Estos textos han sido elegidos específicamente en inglés, porque de este modo se prevé que más cantidad de personas tengan acceso al análisis realizado, motivo por el cual los patrones construidos están también en inglés.

2.3. Metodología de trabajo

La metodología de Trabajo que se ha seguido es la siguiente:

1. **Búsqueda de fuentes documentales disponibles libremente en Internet**

Como ya se ha mencionado anteriormente esta primera fase de la metodología es de suma importancia para el desarrollo del Trabajo y gracias a movimientos como *Open Data*, es posible tener acceso a bases de datos libremente acerca de cualquier tema. Las que resultan de mayor interés son aquellas que tienen asociados textos periodísticos descriptivos o explicativos de los datos y guardan relación con el tema elegido.

2. **Búsqueda de textos periodísticos**

Los textos periodísticos ayudan a conocer cuáles son las frases de más común uso para la interpretación de los datos numéricos.

Del mismo modo que sucede en la búsqueda de bases de datos, lo más importante es poder asociar esos textos a los datos que encontrados, es decir, los puntos uno y dos de la metodología se desarrollan paralelamente.

3. **Análisis de textos**

Una vez recogida toda la información necesaria, se analizan los textos identificando cuales son las relaciones que más están presentes haciendo uso de la herramienta grafica de análisis y las tablas de relaciones de la RST, además de seleccionar frases para la construcción de patrones.

El proceso de selección de frases se explica con más detalle en el apartado *2.5.1 Proceso de selección de frases.*

4. **Construcción y clasificación de patrones**

Después de obtener las frases se construyen y clasifican los patrones por categorías, con las cuales se pretende una mejor compresión del conjunto. Este punto es descrito con más detalle en el apartado *2.6. Construcción de patrones.*

Entre los puntos uno y cuatro se realizan N iteraciones en función de la información que se va obteniendo.

5. **Evaluación del modelo**

Es el punto final de la metodología del trabajo, pero no por ello el menos importante, luego de realizar el análisis y construir los patrones, se realizan las métricas para poder establecer conclusiones.

2.4. Análisis de textos aplicando RST

Haciendo uso de los conocimientos obtenidos al estudiar la Teoría de la Estructura Retórica (*RST*), se procede a analizar los diversos textos seleccionados, Las referencias de los mismos pueden ser consultadas en el apartado *5.3. Referencias de textos* del presente Trabajo.

El objetivo principal del análisis es conocer las relaciones que hay entre las frases o las distintas partes del texto, específicamente en textos periodísticos relacionados con el tema elegido. De modo que se puedan establecer conclusiones fundamentadas.

A continuación se presenta un ejemplo de un texto analizado, haciendo uso de la herramienta *RSTTool* [O'Donnell, 2006]. La cual facilita únicamente el diseño del árbol de relaciones.

Texto 0

Enterprise social networking market heats up: Ovum

Enterprise social networking is "the new battleground for all enterprise collaboration vendors," according to Ovum analyst Richard Edwards.

Ovum estimated the current value of the market exceeds US$500 million and that 10 per cent of organizations in established IT markets have enterprise social networking services.

"Merger and acquisition activity has increased markedly in the past few months, and this has led to new entrants appearing on the enterprise collaboration landscape," Edwards said.

Jive and Yammer are getting the most looks from organizations considering social networking, "but other vendors are generating significant business and revenues from their offerings," he said. Microsoft recently acquired Yammer for US$1.2 billion.

Meanwhile, Salesforce.com recently launched a social media pilot program, while Google recently announced an attempt to bring social media to the enterprise.

"As the business case for investment in enterprise social networking solutions has yet to be proven to business skeptics, some vendors are encouraging independent user adoption in the hope it will prove business value," Edwards said.

"Ovum believes that the business potential offered by enterprise social networking will only be unlocked when necessity dictates a business change."

Como se puede observar en la figura 8, se ha realizado el análisis del texto aplicando la Teoría de la Estructura Retórica, en ella se pueden observar las relaciones existentes entre las distintas partes del texto.

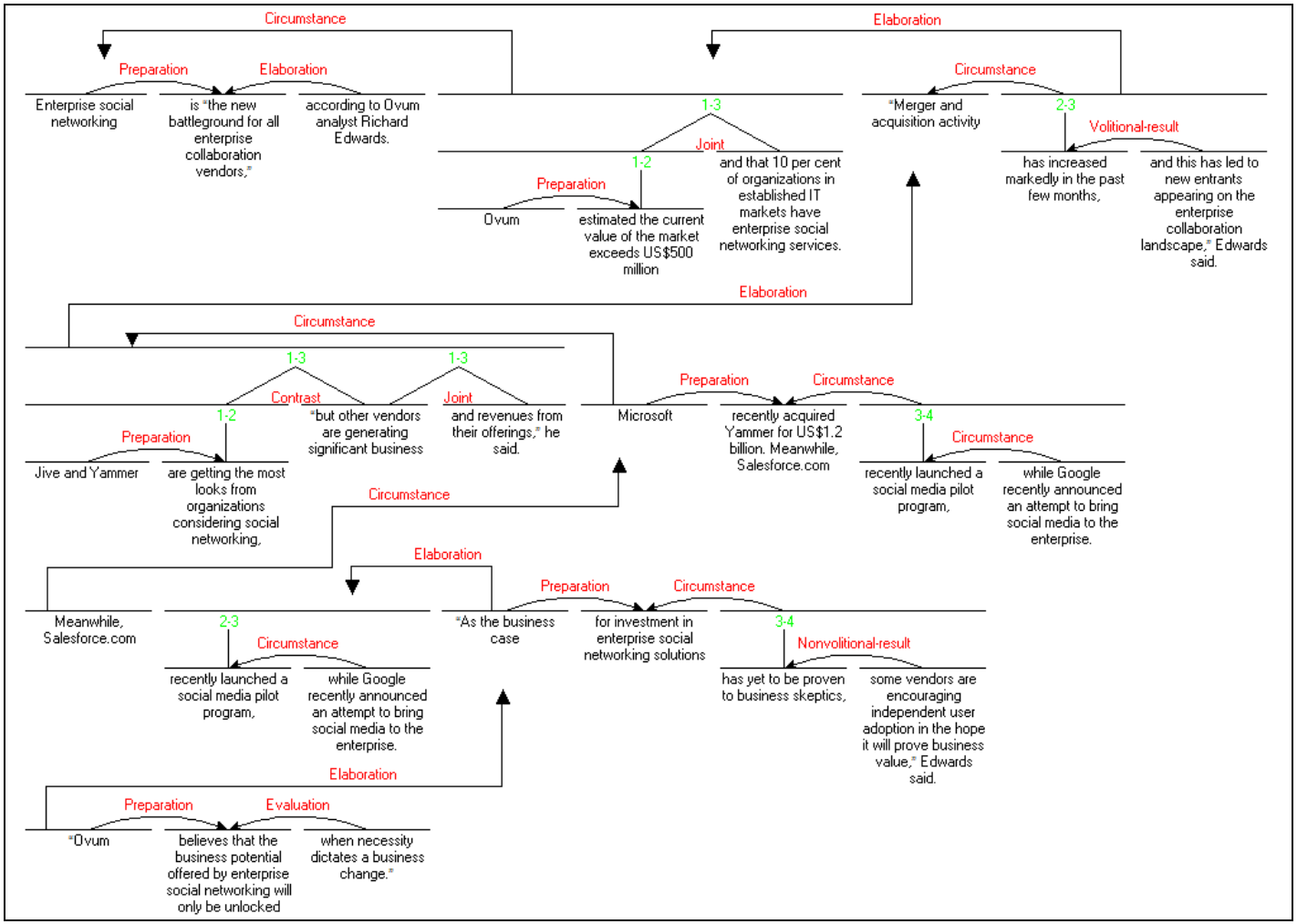


Figura 8: *Análisis de textos aplicando RST.*

En la tabla 4, se puede ver el resultado del análisis, en el que se aprecia que en su gran mayoría las relaciones presentes son circunstanciales ya que hacen referencia a un marco contextual, situacional o temporal, por el contrario de las relaciones de resultado de causa voluntaria o involuntaria, de contraste y de evaluación que sólo aparecen una única vez.

Relación	Nº	Media
Circumstance	8	32%
Contrast	1	4%
Elaboration	5	20%
Evaluation	1	4%
Joint	2	8%
Preparation	6	24%
Volitional - result	1	4%
Non – Volitional result	1	4%

Tabla 4: *Resultados del análisis.*

Este tipo de medidas resultan de gran ayuda para formar una idea de cuáles son los pares de relaciones del tipo de textos del tema seleccionado.

Suponiendo que fuese el resultado final después de analizar gran cantidad de textos, hipotéticamente se puede concluir que: en los artículos periodísticos relacionados con el comportamiento humano y el uso que hacen las personas de las redes sociales el 32% de los textos transmiten una idea dependiendo del marco conceptual o temporal, muy seguidamente con un 24% las relaciones de preparación anticipan al lector la idea que va a ser transmitida, con un 20% de relaciones de elaboración aportan más información a la idea principal. Y de ese modo se continúa estableciendo conclusiones a partir de los resultados.

Es importante tener en cuenta, que la RST es estándar de *facto*, es decir, que se utiliza gracias a la utilidad que presenta en el análisis de textos, pero no es totalmente exacta por lo que no todos los análisis de un mismo texto son cien por ciento iguales, ya que todas las personas no poseen los mismos criterios, es decir, el factor humano interviene notablemente.

Para ver otros análisis basados en la RST y haciendo uso de la herramienta *RSTTool* ir al apartado *5. Anexos*.

2.5. Construcción del modelo

2.5.1. Proceso de selección de frases.

El proceso o metodología de selección de frases está conformado por las siguientes etapas:

a. **Identificación de variables:**
   - **Numéricas [A]:** son adjetivos numerales que pueden ser encontrados en los textos expresados numéricamente (1, 2, 3) o escritos (uno, dos, tres). También comprende los números ordinales y cardinales.
   - **Porcentuales [B]:** son adjetivos numerales pero con una restricción, deben estar seguidos de "%" o "por ciento".
   - **De lugar [C]:** son sustantivos que identifican sitios o lugares específicos como pueden ser: Madrid, casa, museo, entre otros.
   - **De tiempo [D]:** son adverbios de tiempo que hacen referencia a algún periodo específico, por ejemplo: hoy, ayer, mañana, hace un mes, diez años atrás, entre otros.
   - **De fecha [E]:** variables que pueden ser representadas por el conjunto día, mes y año o individualmente.
   - **De duración [F]:** son variables que indican la duración de alguna actividad, normalmente irán seguidos por horas, minutos y/o segundos.
   - **De Grupo [H]:** variables utilizadas para identificar individuos, si se trata de hombres, mujeres, niños, adolescentes, y personas mayores (siempre haciendo referencia a más de una persona).
   - **Calificativas [I]:** son adjetivos calificativos que aportan más características del individuo u objeto al que se hace referencia.
   - **Monetaria [K]:** se refiere a aquellas variables numéricas que poseen algún símbolo como $, €, y van seguidas de palabras como millones, billones, miles, entre otros.
   - **Actividad [L]:** variables que hacen referencia a la realización de alguna actividad en concreto.
   - **Organización [M]:** son nombres propios de organizaciones de las cuales se cita algún estudio realizado.
   - **Objeto [N]:** este grupo de variables pertenecen todas aquellas que no forman parte de los anteriormente descritos.
   - **Redes [O]:** es el conjunto de variables que representa el nombre de cada una de las redes sociales incluyendo el Internet.

b. **Búsqueda de frases:** son tomadas en cuenta las oraciones que aparezcan con más periodicidad en la mayor parte de los artículos seleccionados o en un mismo artículo, en total 20 textos y que contengan al menos dos de las variables mencionadas en el punto anterior.

También es importante tener en cuenta el grado de discrepancia que pueden tener las frases que aparecen más comúnmente en los textos, es decir, pocas veces se van a encontrar dos o más oraciones exactamente iguales, y no por eso son descartadas,

siempre y cuando presenten el mismo patrón o trasmitan el mismo mensaje. En la figura 9, se muestra un ejemplo del grado de discrepancia que puede haber en las frases, que aunque no son exactamente iguales siguen la misma estructura y transmiten la misma idea, es decir, las partes subrayadas antes y después de los verbos seleccionados en azul, son variables del mismo tipo.

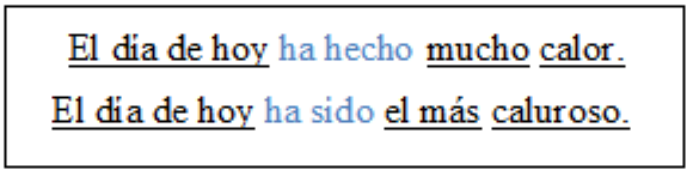


Figura 9: *Grado de discrepancia entre frases.*

2.6. Construcción de patrones

A medida que se eligen las frases que siguen el proceso de selección descrito en el punto anterior, se van creando patrones que pueden o no ser fieles al número total de variables de las frases originales.

2.6.1. Características de los patrones

Para cada patrón que conforma el modelo, se proporciona la siguiente información:

2.6.1.1. Category

Los patrones forman parte de categorías dependiendo de las características que poseen. Ver referencia en la tabla 1 y tabla 2 del punto *1.3.1.1. Relaciones* de éste documento.

2.6.1.2. Reference

Se utiliza la notación Rx, donde "R" significa referencia y "x" es reemplazado por números enteros positivos, este campo tiene la finalidad de señalar las fuentes de las cuales son obtenidas las frases originales.

2.6.1.3. Original Sentence

Es la frase original de más común uso en textos periodísticos elegida siguiendo el proceso de selección de frases, a partir de la cual se genera el patrón.

2.6.1.4. Template

Son de primer o segundo orden, o sucesivos, dependiendo de la cantidad de patrones que se obtienen a partir de la frase original. Normalmente el *template* de primer orden es específico del tema elegido y el de segundo orden es más generalizado.

En la figura 10, se puede observar el formato en el que se agrupa toda la información anteriormente descrita.

Category:
Reference:
Original Sentence:
Template:

Figura 10: *Tabla informativa de patrones.*

2.6.2. Patrones obtenidos

Para hacer más fácil la comprensión de cuáles son las partes de la frase que se mantienen fijas y cuáles se sustituyen, las variables que están escritas entre corchetes y coloreadas de azul representan las que se reemplazan.

A continuación se puede observar uno de los patrones que forma parte del modelo.

Category:	**Contrast**
Reference:	R2
Original Sentence:	[Women] have been significantly [more] likely to use [social networking sites] than [men] since 2009.
Template 1st order:	[H] have been [*1] likely to use social networking sites than [H].
Template 2nd order	[H] have been [*1] likely to use [N] than [H].
*1. {more, less}	

El patrón en cuestión pertenece a la categoría contraste, ya que realiza una comparación basándose en el género de las personas, para saber cuan propensos son a utilizar las redes sociales.

De la frase original se obtienen dos *templates*, el de primer orden está compuesto por 3 variables, dos de las cuales son del tipo grupo y la tercera es comparativa. El de segundo orden tiene una variable más que el anterior y es de tipo objeto.

Utilizando esos patrones se pueden obtener los siguientes ejemplos de frases:

- Men have been more likely to use social networking sites than women.
- Children have been less likely to use social networking sites than teens.
- Adults have been more likely to use mobile phones than children.
- Women have been less likely to use glasses than men.

Para consultar los demás patrones que forman el conjunto complete, ir al apartado *5.1 Conjunto de patrones.*

Capítulo 3

EVALUACIÓN

3. EVALUACIÓN

En esta sección se presenta la evaluación realizada tanto para el análisis de los textos como para el conjunto de 101 patrones.

Para la evaluación del análisis de textos, se establecen métricas que sirven de ayuda para realizar comparaciones entre los textos relacionados con el tema elegido y demás textos periodísticos, con la finalidad de conocer las características propias y las diferencias con respecto a los demás.

En cuanto a la evaluación de los patrones, se listan una serie de características o datos de interés y se establecen ciertas métricas que sirven para entender mejor cada uno de los patrones del conjunto total.

Finalmente, se presentan conclusiones del resultado de evaluar ambas fases principales análisis realizado y un ejemplo de texto representativo de textos explicativos de datos numéricos.

3.1. Métricas de textos

De los 20 textos periodísticos utilizados para hacer el análisis, se han estudiado el 30% del total, seleccionando segmentos estratégicos debido a lo extensos de los mismos, para poder medir numéricamente las tendencias de las relaciones o pares de relaciones presentes.

A continuación en la tabla 5, se muestra un resumen de los resultados obtenidos.

Relación	Nº	Media
Circumstance	23	11.61%
Contrast	28	14.14%
Elaboration	69	34.84%
Evaluation	3	1.51%
Evidence	5	2.52%
Joint	15	7.57%
List	23	11.61%
Preparation	23	11.61%
Summary	1	0.50%
Otherwise	3	1.51%
Volitional - result	4	2.02%
Non – Volitional result	1	0.50%

Tabla 5: *Resultados generales obtenidos de análisis de textos.*

Como se puede observar doce de las veintisiete relaciones tanto nucleares como multinicleares definidas en la Teoría de la Estructura Retórica están presentes en textos periodísticos relacionados con las redes sociales y el uso que hacen las personas de ellas. De las cuales *joint, list y contrast* son multinucleares.

Aunque en los segmentos de textos seleccionados no aparezcan las demás relaciones, no implica que no puedan formar parte de este tipo de artículos, sino que es posible que no se manifiesten con tanta periodicidad.

Es importante tener en cuenta que, a pesar de que la Teoría de la Estructura Retorica es utilizada y validada hoy en día para el análisis de textos, no es cien por cien precisa, ya que interviene el factor humano, es decir, un mismo texto examinado por un número *X* de individuos puede o no dar el mismo tipo o cantidad de relaciones, porque no todos razonan o piensan de la misma manera.

Al tratarse de textos que parten de datos numéricos obtenidos de bases de datos, es perceptible:

- La necesidad de elaborar un texto que explique detalladamente el significado del dato.
- La certeza con la que se afirma o no cierta situación.
- El uso de los datos dentro del texto para hacer más sencilla la comprensión del mismo.
- La mezcla en algunos casos de tecnicismos con un lenguaje cotidiano y fácil de entender.

Con respecto a los pares de relaciones más comunes encontrados en textos periodísticos, en la tabla 6 se muestra un resumen del porcentaje de la presencia de cada par dentro de los textos analizados.

Pares de relaciones	Nº	Media
Circumstance – Contrast	7	11.29%
Elaboration – Contrast	13	21.66%
Elaboration – Evidence	6	9.67%
Elaboration – List	6	9.67%
Elaboration – Otherwise	3	4.83%
Elaboration – Volitional Result	4	6.45%
Preparation – Elaboration	15	24.19%
Preparation – Circumstance	8	12.90%

Tabla 6: *Resultados generales de análisis de textos.*

Los pares de relaciones que más presencia tienen en los textos son *Preparation – Elaboration* y *Elaboration – Contrast* con una media de aparición del 24.19 % y 21.66% respectivamente.

Dichos resultados son de esperarse en este tipo de textos ya que los números hacen que sea muy fácil establecer comparaciones entre dos ideas elaboradas. De hecho este par de relaciones es un buen conjunto para comenzar la redacción de un texto periodístico del tema elegido.

Al comparar este tipo de texto, se encontró que guarda más vinculación con **textos económicos**, ya que surgen a partir de datos numéricos reales que son explicados con la finalidad de que la mayoría de las personas sean capaces de comprenderlas utilizando ciertos tecnicismos mezclados con un lenguaje de común uso.

Por el contrario de la prensa sensacionalista o amarillista, de las cuales no cabe la menor duda que se basen en hechos reales y tengan evidencia de los mismos, pero no se ciñen únicamente a la información que puede ser cotejada sino que aportan un poco de dramatismo para captar la atención del lector.

3.2. Métricas de patrones

En líneas generales la información más representativa del conjunto formado por 101 patrones es la siguiente:

- 8% son patrones de contraste.
- 6% son patrones mixtos de contraste y circunstanciales.
- 44% son patrones de evidencia.
- 2% son patrones mixtos de evidencia y circunstanciales.
- 27% son patrones circunstanciales.
- 5% son patrones de justificación.
- 1% son patrones de evaluación.
- 2% son patrones de resultado voluntario.
- 6% son patrones de elaboración.

En la figura 11, se muestra una visión más gráfica de los datos.

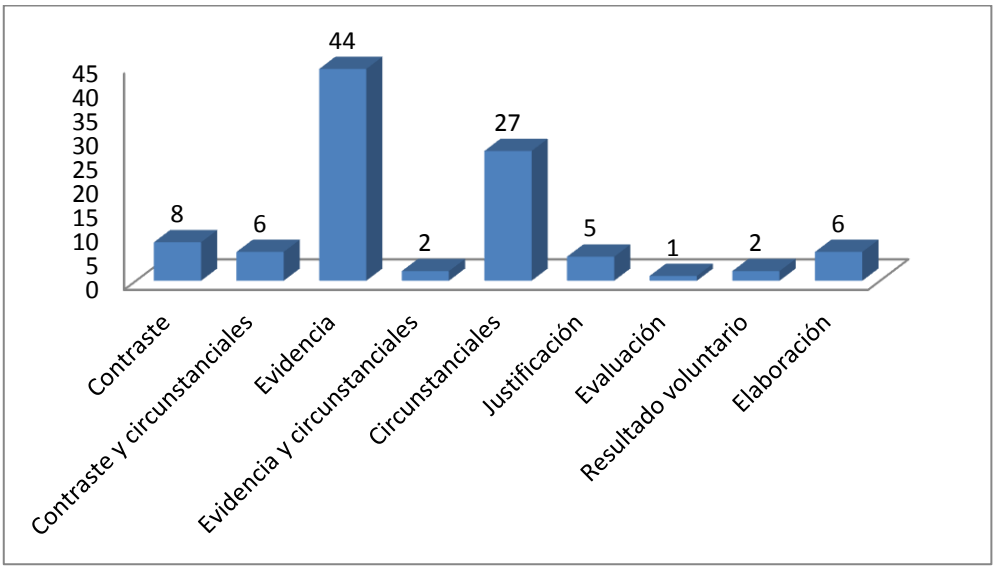


Figura 11: *Ocurrencias de patrones usando RST*

Otra información relevante:

- 62% de los patrones presentan o hacen referencia a valores porcentuales.
- 5% de los patrones hacen referencia a estudios realizados por empresas u otros individuos.
- A lo sumo cada patrón contiene cinco variables para ser sustituidas.
- El lenguaje utilizado en cada patrón es de común uso, por lo que son de fácil comprensión.
- La construcción de patrones está basada en el estudio realizado sobre los textos teniendo en cuenta las relaciones de más común uso, de modo que es posible la creación de textos que elaboren y expliquen ideas claramente.

Puede resultar curioso que el número de patrones creados no se corresponda con el número de relaciones que se obtienen después de analizar los textos (por ejemplo: 28% de relaciones de contraste en el análisis de textos, 14% de patrones de contraste), esto se debe a que las frases seleccionadas para la construcción de los patrones en muchos casos no forma parte de los fragmentos de textos analizados.

La razón por la cual no se eligieron exactamente los mismos tramos, fue para comprobar si existía la presencia de alguna otra relación que no había sido tomada aun en cuenta.

Sin embargo como es de esperarse, destacan los patrones que hacen uso de la presentación de evidencias, lo que se considera lógico en un análisis de datos numéricos. Seguidamente se encuentran los patrones circunstanciales que suelen estar presentes siempre que se desea hacer referencia a marcos temporales o situacionales.

3.3. Conclusiones de la evaluación

Una vez establecidas las métricas y comentado los resultados, se llega a la conclusión que las frases más habituales están orientadas a sustentar con evidencias cuantitativas las afirmaciones realizadas (se presenta la relación *Elaboration* en la mayor parte con un 34.84%). Se han identificado el mayor número de patrones con este tipo de frase (44 patrones de evidencia). Por lo que el par de relaciones *Elaboration – Evidence* se encuentra presente en los textos con un 9.67%.

Las relaciones de contraste y circunstanciales tienen una alta presencia (especialmente tras la relación *Elaboration*). Estas relaciones son muy útiles para permitir al lector valorar las medidas cuantitativas. También se han identificado un número de patrones para estas relaciones (35 en total).

Por su parte, las relaciones multinucleares *List* y *Join* son también muy utilizadas para enumerar los diferentes aspectos de los datos analizados, mostrando en forma de enumeración las características que explican las medidas.

La relación *Preparation* tiene también una importante presencia de 11.61%, lo cual es de esperar cuando se trata de anticipar y contextualizar la presentación de valores cuantitativos.

El par de relación *Elaboration – Volitional Result,* tiene una presencia de 6.45% en los textos explicativos de datos numéricos cuando se quiere expresar la consecuencia o efecto de una acción. Para estas relaciones se ha identificado un total de 13 patrones.

En los artículos periodísticos relacionados con el comportamiento humano y el uso que hacen las personas de las redes sociales, el 11.61% anticipan al lector la idea que va a ser transmitida, el 34.84% de los textos elaboran la idea mediante marcos conceptuales o temporales, estableciendo comparaciones o listando una serie de características (11.61% de los textos).

3.4. Perfil de texto explicativo representativo

Luego de analizar los textos periodísticos y construir patrones, se puede generar un perfil de texto explicativo representativo, compuesto por las siguientes relaciones o pares de relaciones:

a. Preparation
b. Elaboration
    a. Circumstance
    b. List
    c. Joint
    d. Contrast
c. Volitional Results

Lo que no implica que deba cumplirse estrictamente, puede contener dichas relaciones o más pero con un orden aleatorio.

Un ejemplo de texto que cumpla con la estructura planteada es el siguiente:

10 striking conclusions of the Social Media around the World 2012 study

Today InSites Consulting published the third edition of its study "Social Media around the world" in collaboration with data and sampling partner SSI and translation agency No Problem!. More than 7,800 respondents from nineteen different countries took part in the new survey and this article presents the study's ten most striking conclusions.

1. There are more than 1.5 billion social network users worldwide.
2. Fast adoption of smartphones boosts social media use.
3. Most internauts use no more than two social network sites.
4. Pinterest and Instagram are the rising stars.
5. Klout is a niche.
6. Half of consumers are connected to at least one brand
7. 1 in 2 consumers occasionally post brand-related content.
8. Pinterest probably more interesting for brands than Instagram.
9. People don't really trust brand fans.
10. 80% of people are open to co-creation

The main conclusion is that us social media fanatics should take care to remain firmly anchored in reality. Not everyone wants to try every new platform and not all of us are waiting for some obscure software update. On the contrary, the reality is that the average consumer has more or less shaped the social media landscape and structural changes are unlikely in the next few years. Consumers are active on just one or two social networks. They have integrated these sites into their everyday lives and while mobile technology is speeding up this process, this is as far as it goes for now. A fringe minority of consumers juggles more than four social network accounts or attaches at least some importance to Klout.

Of course, there is nothing wrong with being a social media fanatic, but when advising companies we need to keep our feet on the ground and be realistic.

En la figura 12, se puede observar el análisis del texto aplicando RST. Como se puede observar el texto contiene todas las relaciones principales del tipo de texto periodístico relacionado con el tema elegido.

Además de estar presentes los pares de relaciones:

- *Preparation – Elaboration*
- *Elaboration – List*
- *Elaboration – Volitional Result*
- *Elaboration – Contrast*

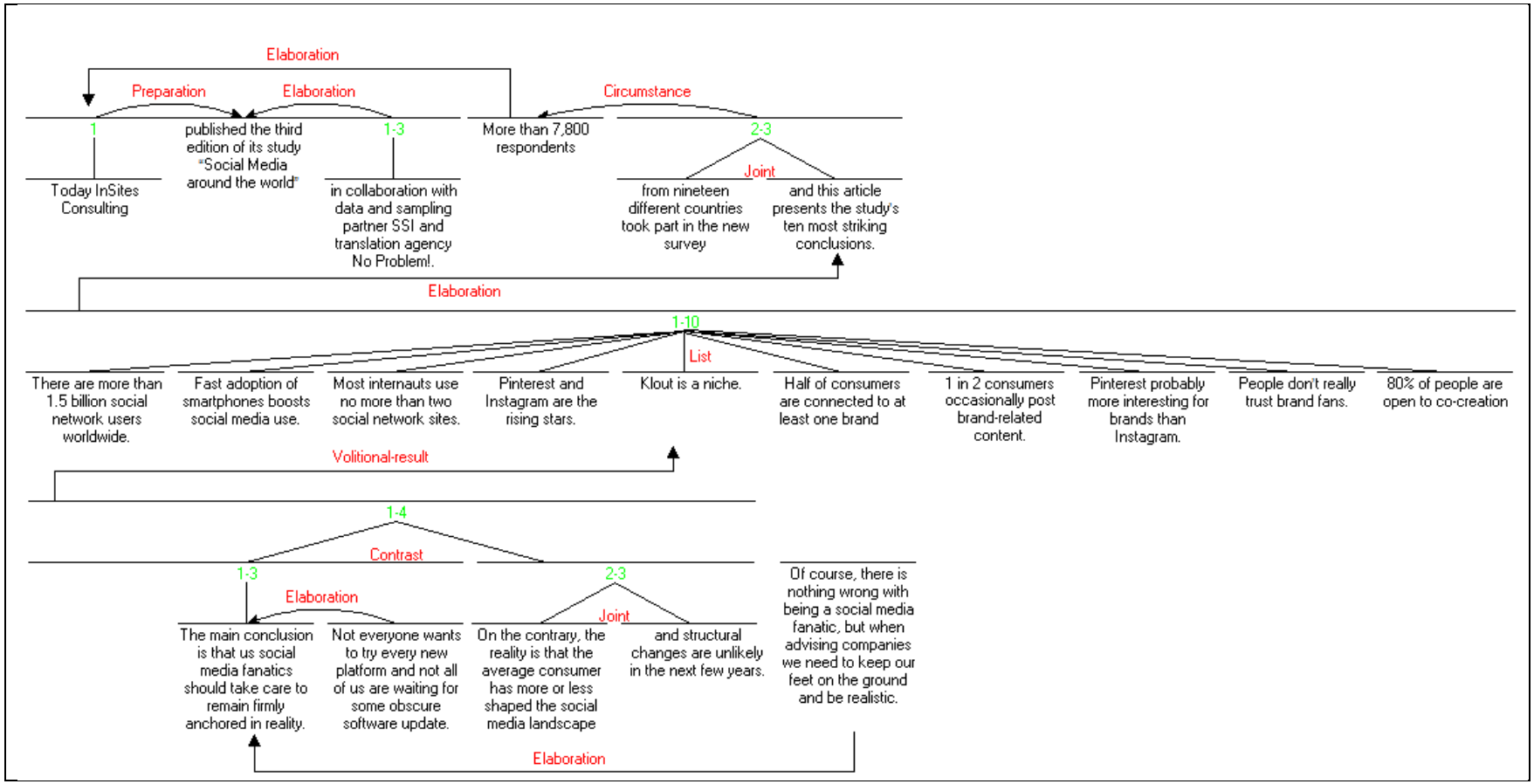


Figura 12: *Análisis de perfil de texto explicativo representativo aplicando RST.*

Capítulo 4

CONCLUSIONES Y FUTURAS LÍNEAS DE TRABAJO

4. CONCLUSIONES

El análisis de textos explicativos del significado de datos cuantitativos y la construcción de un modelo, conformado por un total de 101 patrones, ha sido llevado a cabo siguiendo una metodología definida a lo largo del desarrollo del Trabajo.

La cual estaba formada por un total de 5 pasos que engloban la búsqueda de fuentes documentales y bases de datos que estuvieran relacionadas entre sí y además con el tema elegido el comportamiento humano y el uso que hacen las personas de las redes sociales, el análisis de textos aplicando la Teoría de la Estructura Retórica lo cual añade más valor al Trabajo por ser una teoría utilizada y validada en la actualidad, y la construcción de los patrones haciendo uso de frases elegidas de los textos analizados.

Para estos cuatro primeros pasos se podían realizar un número N de iteraciones dependiendo de la información recogida y del grado de detalle que se quisiera dar tanto al análisis como a la construcción de los patrones.

El último paso de la metodología comprendía la evaluación del modelo, en la cual se establecieron métricas para poder hacer comparaciones entre las características de los textos del tema elegido y las características de otros textos de cualquier tema.

También para poder establecer vínculos entre las relaciones de más común uso en los textos y el conjunto de patrones creados, aunque en muchos casos, el número de relaciones no se asemejaba al de los patrones. La razón de ello es que fueron seleccionados fragmentos distintos para intentar encontrar más relaciones o patrones que no habían sido tomados en cuenta hasta el momento.

No hay que olvidar que un factor importante en el uso de la RST, es el factor humano y que un mismo texto puede ser analizado de distintas maneras y arrojar resultados diferentes pero no por ello es incorrecto, todo lo contrario, pueden ser soluciones alternativas a un mismo problema.

En cuanto a los objetivos que fueron propuestos desde el inicio del Trabajo, se han superado satisfactoriamente, es decir, se comprendió que uno de los grandes objetivos de la Inteligencia Artificial es intentar simular el razonamiento humano para poder generar Lenguaje Natural y que a lo largo del tiempo se han desarrollado métodos y/o programas que mantienen conversaciones entre persona – máquina de un modo tan fluido como si se tratase de dos personas, además de sistemas de datos a textos que a través de números y una base de conocimientos son capaces de generar textos explicativos que son de gran ayuda para los seres humanos.

Con respecto a la búsqueda de fuentes documentales y textos periodísticos, se ha hecho una indagación exhaustiva en distintos sitios webs basados en movimientos como *Open*

Data, los cuales publican bases de datos que son de libre acceso a todas las personas y que pueden ser utilizadas sin tener problemas de patentes o *copyright.* Se puede decir que el objetivo más difícil de superar fue este, ya que había que encontrar bases de datos tuviesen vinculados textos explicativos de los datos o viceversa.

El proceso de análisis sin duda alguna, fue el que se llevó más tiempo de desarrollar, por la implicación de la Teoría de la Estructura Retorica, la cual era necesaria estudiar y comprender para poder hacer el análisis de los textos, además del tiempo utilizado para familiarizarse con el uso de la herramienta *RSTTool,* pero una vez realizado el primer análisis los demás fueron elaborados de una manera más fluida.

Los patrones también tuvieron su grado de complejidad, porque aunque el resultado final muestra que son simples y concretos, fue necesario un aproximado de cuatro o cinco revisiones para depurarlos todos.

El objetivo final del Trabajo era llevar a cabo una evaluación en la que se demostrara la importancia de las relaciones presentes en los textos que hacen que sea específico de un tema particular. Además de vincular también las relaciones o pares de relaciones encontradas con los patrones creados.

En general, el Trabajo ha sido llevado a cabo en el tiempo que ha sido estipulado desde un comienzo, por lo cual aparte de los objetivos propios del Trabajo se han cumplido los indirectos como lo son: organización efectiva del tiempo, conocimiento de cómo gestionar un proyecto que depende únicamente del autor del mismo y ser capaz de poner en práctica los conocimientos adquiridos durante el periodo de estudios.

Concluyendo, se ha podido obtener las características principales de los textos periodísticos relacionados con el tema elegido y establecer similitudes y/o diferencias con respecto a otros textos, se identificaron las relaciones o pares de principales presentes que se consideran indispensables para la correcta y coherente construcción del texto y se crearon un total de 101 patrones partiendo de frases claves seleccionadas meticulosamente de los textos analizados siguiendo un proceso compuesto de varios pasos definido precisamente para ello.

Se consideran superados los objetivos del análisis de textos explicativos de datos cuantitativos, el cual puede ser utilizado confiando en que la información que es representada, es certera y fundamentada en teorías.

Por lo tanto, se estima que el Trabajo resulte de utilidad independientemente de la línea futura de trabajo que las tres que se propondrán en el siguiente apartado.

4.1. Futuras líneas de trabajo

En este apartado se describen las posibles líneas de Trabajo a tener en cuenta para la continuación del Proyecto en el futuro.

- Ampliar el conjunto de patrones.

El conjunto actual está conformado por un total de 101 patrones, un número bueno para una base de conocimientos de un sistema generador de textos, pero mientras más patrones se obtengan será mucho mejor. Para ello se recomienda se sigan los pasos detallados el apartado *2.5.1. Proceso de selección de frases* para identificar las frases que servirán como base en la construcción de otros patrones y consultar el apartado *2.6. Construcción de patrones* para saber la información que es necesaria obtener de cada uno, como lo es la categorización, la referencia del texto del cual fue obtenido y la frase original que se utiliza como ejemplo.

- Ampliar el estudio y extrapolarlo a otros temas.

Cuando se habla de ampliar el estudio ya realizado se refiere a seguir obteniendo más textos relacionados con el tema del comportamiento humano y el uso que hacen las personas de las redes sociales y analizarlos haciendo uso de la Teoría de la Estructura Retórica, para lo cual será necesario consultar el apartado *1.3. Teoría de la Estructura Retórica* y sus correspondientes sub apartados *1.3.1. Definiciones de Estructura, Esquemas y Relaciones* y *1.3.2. Relación con la Lingüística Computacional*, de modo que se pueda comprender como hacer uso de la misma y se proceda a hacer el análisis correspondiente.

Es una línea de trabajo futura recomendable, pero incluso lo sería aún más si la información ya obtenida sirviera como punto de partida para hacer comparaciones con otros temas los cuales ayudaría a tener una idea sobre las diferencias o similitudes entre unos textos y otros.

- Construcción de un Sistema Generador de textos automáticos.

Esta opción es viable e incluso la más ambiciosa, ya que fue la motivación principal por la cual realizar el análisis de los textos haciendo uso de la Teoría de la Estructura Retórica, que ha ayudado a comprender las relaciones de más común uso en el tipo de textos utilizados y la realización de patrones relacionados con el mismo tema, hacen que este Trabajo de Fin de Grado pueda ser utilizado en la construcción de un Sistema Generador de textos automáticos. El estudio realizado formaría parte del modelo del sistema, además los patrones realizados constituirían la base de conocimientos que se utilizaría para la generación de textos.

Capítulo 5

ANEXOS

5. ANEXOS

5.1. Conjunto de Patrones

5.1.1. De contraste

Category:	Contrast
Reference:	R9
Original Sentence:	[Internet] users are [more] likely than non-users to receive help from [core network members].
Template 1st order:	[O] [*2] are [*1] likely than [*2] to [L].
Template 2nd order	[H] are [*1] likely than [H] to [L].
*1. {more, less}	
*2. {users, non-users}	

Category:	Contrast
Reference:	R20
Original Sentence:	Fully [65%] of [adult] internet users [now] say they use [a social networking site] like [MySpace], [Facebook] or [LinkedIn], up from [61%] [one year ago].
Template 1st order:	Fully [B] per cent of [H] internet users [D] say they use [a social networking site] like [O], [O] or [O], up from [B] per cent [D].
Template 2nd order:	Fully [B] per cent of [H] say they use [N], up from [B] per cent [D].

Category:	Contrast
Reference:	R15
Original Sentence:	[More] than [50%] clamming to have [more] than [100] contacts in their network.
Template 1st order:	[*1] than [B] per cent claiming to have [*1] than [A] contacts in their network.
Template 2nd order:	[*1] than [B] per cent claiming to have [N].
*1. {more, less}	

Category:	Contrast and Circumstance
Reference:	R2
Original Sentence:	In [December 2012], [71%] of [women] were users of [social networking sites], compared with [62%] of [men].
Template 1st order:	In [E], [B] per cent of [H] were users of social networking sites, compared with [B] per cent of [H].
Template 2nd order	In [E], [B] per cent of [H] were users of [N], compared with [B] per cent of [H].

Category:	Contrast and Circumstance
Reference:	R2
Original Sentence:	Between [February 2005] and [August 2006], the use of [social networking sites] among young adult internet users ages [18-29] jumped from [9%] to [49%].
Template 1st order:	Between [E] and [E], the use of [social networking sites] jumped from [B] per cent to [B] per cent.
Template 2nd order:	Between [E] and [E], the use of [N] moved [B] per cent to [B] per cent.

Category:	Contrast and Circumstance
Reference:	R19
Original Sentence:	Only [6%] of [Facebook] users use this platform [more] than [once per month].
Template 1st order:	Only [B] percent of [O] users use this platform [*1] than [D].
Template 2nd order:	Only [B] percent of [H] use [N] [*1] than [D].
*1. {more, less}	

5.1.2. De evidencia

Category:	Evidence
Reference:	R8
Original Sentence:	Some [37%] of [internet] users aged [18-29] use blogs or social networking sites as a venue for [political or civic involvement], compared to [17%] of online [30]-[49] year olds, [12%] of [50]- [64] year olds and [10%] of internet users over [65].
Template 1st order:	Some [B] of [O] users aged [A] use social networking sites as a venue for [L].
Template 2nd order	Some [B] of [H] aged [A] use [N] as a venue for [L].

Category:	Evidence
Reference:	R1
Original Sentence:	[10] per cent of [organizations in established IT markets] have [enterprise social networking service].
Template 1st order:	[B] per cent of [H] have [O] [N].
Template 2nd order:	[B] per cent of [H] have [N].

Category:	Evidence
Reference:	R4
Original Sentence:	[28%] of [nine and ten-year-olds] in [UK] use [social networking websites]
Template 1st order:	[B] per cent of [H] in [C] use social networking websites.
Template 2nd order:	[B] per cent of [H] in [C] use [N].

Category:	Evidence
Reference:	R4
Original Sentence:	About [28%] of [UK] [9-10] year olds and [59%] of [11]-[12] year olds operate [a social networking profile].
Template 1st order:	About [B] per cent of [H] operate a social networking profile.
Template 2nd order:	About [B] per cent of [H] operate [N].

Category:	Evidence
Reference:	R5
Original Sentence:	[23] per cent of [Facebook's] users check their [account] [5] or more times [daily].
Template 1st order:	[B] per cent of [O] users check their account [A] or [*1] times [D].
Template 2nd order:	[B] per cent of [H] check their [N] [A] or [*1] times [D].
*1. {more, less}	

Category:	Evidence
Reference:	R5
Original Sentence:	[Facebook] users get [more] [social support] than [other people].
Template 1st order:	[O] users get [*1] social support than other people.
Template 2nd order:	[H] get *1 [N] than [H].
*1. {more, less}	

Category:	Evidence
Reference:	R6
Original Sentence:	And social media users are even [more] likely to be active: [82%] of social network users and [85%] of [Twitter] users are group participants.
Template 1st order:	[B] per cent of social network users are participants of [O].
Template 2nd order:	[B] per cent of [H] are participants of [O].

Category:	Evidence
Reference:	R6
Original Sentence:	[68%] of all [Americans (internet users and non-users alike)] said [the internet] has had [a major] impact on the ability of [groups] to communicate with [members].
Template 1st order:	[B] per cent of all [H] said social network has had [*1] impact on [L] [N].
Template 2nd order:	[B] per cent of all [H] said [N].
*1. {a major, a minor}	

Category:	Evidence
Reference:	R7
Original Sentence:	Nearly [one] in [ten] social network users (8%) joined an online group focused on [community issues] in the preceding [twelve months].
Template 1st order:	Nearly [A] in [A] social network users joined an online group focused on [N] in the preceding [D].
Template 2nd order:	Nearly [A] in [A] [H] joined [O] in the preceding [D].
Template 2nd order (variation):	Nearly [A] in [A] [H] joined [N] in the preceding [D].
Template 3rd order:	Nearly [B] per cent of [H] joined [N] in the preceding [D].

Category:	Evidence
Reference:	R13
Original Sentence:	Over [1 Billion] [People] Use [Social Networks Today].
Template 1st order:	Over [A] [H] Use [O] [D].
Template 1st order (variation):	Over [A] [H] Use [N] [D].
Template 2nd order:	Over [A] [H] use [N].

Category:	Evidence
Reference:	R12
Original Sentence:	[40] per cent of [people] have confessed to looking at [their partner's private messages and emails].
Template 1st order:	[B] per cent of [H] have confessed to looking at [N].

Category:	Evidence
Reference:	R9
Original Sentence:	[The Social Ties survey] asked about two types of connections [people] have in [their social networks].
Template 1st order:	[M] asked about [N] that [H] have in [O].

Category:	Evidence
Reference:	R10
Original Sentence:	Suddenly, [people] in [their 30s and 40s] were using [social networks] to find [long lost friends].
Template 1st order:	Suddenly, [H] were using social networks to find [H].
Template 2nd order:	Suddenly, [H] were using [O] to find [N].
Template 3rd order:	[H] were using [N] to find [N].

Category:	Evidence
Reference:	R17
Original Sentence:	[Bloggers] are [72%] [*1] likely to belong to [a local group].
Template 1st order:	[H] are [B] per cent [*1] likely to belong to [M].
Template 2nd order:	[H] are [B] per cent [*1] likely to [L] [N].
*1. {more, less}	

Category:	Evidence
Reference:	R16
Original Sentence:	[3%] of [Facebook] users say they plan to spend [*1] time on [the site] in [the coming year].
Template 1st order:	[B] per cent of [O] users say they plan to spend [*1] time on [N] in [D].
Template 2nd order:	[B] per cent of [H] plan to spend [*1] time on [N] in [D].
*1. {more, less}	

Category:	Evidence
Reference:	R15
Original Sentence:	[Facebook] is used by [more] than [95%] of [social media users].
Template 1st order:	[O] is used by [*1] than [B] per cent of social media users.
Template 2nd order:	[O] is used by [*1] than [B] per cent of [H].
*1. {more, less}	

Category:	Evidence
Reference:	R15
Original Sentence:	On average, [facebook] [users] spend around [18 minutes] on [the site] each time they access it.
Template 1st order:	On average, [H] spend around [F] on [O] each time they access it.
Template 2nd order:	On average, [H] spend around [F] on [N].

Category:	Evidence
Reference:	R16
Original Sentence:	[Two]- [thirds] of online [American] [adults] (67%) are [Facebook] users.
Template 1st order:	[A]- [A] of online [I] [H] are [O] users.
Template 2nd order:	[A]- [A] of [I] [H] are [N] users.
Template 3rd order:	[B] per cent of [H] are [N].

Category:	Evidence
Reference:	R15
Original Sentence:	Around [7] in [10] (70%) use [laptops] to access their [social media sites].
Template 1st order:	Around [A] in [A] use [N] to access their [social media sites].
Template 2nd order:	[H] use [N] to access [O].
Template 3rd order:	Around [B] per cent of [H] use N to access [O].

Category:	Evidence
Reference:	R18
Original Sentence:	Nearly [seven] in [ten] (69%) [teens], ages [12]- [17], have [a computer].
Template 1st order:	Nearly [A] in [A] [H] ages, [A]- [A] have, [N].
Template 2nd order:	[B] per cent of [H] have [N].

Category:	Evidence and Circumstance
Reference:	R12
Original Sentence:	Of [2,400] [American adults] that have admitted [to cheating] at least one [this past year].
Template 1st order:	Of [A] [H] that have admitted to [L].

Category:	Evidence and Circumstance
Reference:	R1
Original Sentence:	[Merger and acquisition activity has increased] markedly in [the past few months].
Template 1st order:	[L] [*1] markedly in [D].
*1. { has increased, has decreased}	

5.1.3. De circunstancia

Category:	Circumstance
Reference:	R3
Original Sentence:	Fully [95%] of all [teens] ages [12]- [17] are [now] [online].
Template 1st order:	[B] per cent of all [H] ages [A]- [A] are [D] [*1].
*1. {online, offline}	

Category:	Circumstance
Reference:	R10
Original Sentence:	[Social networks] are now the [fourth] [*1] popular online category, even ahead of personal e-mail.
Template 1st order:	Social networks are now the [A] [*1] popular online category.
Template 2nd order:	[O] is now the [A] [*1] popular online category.
Template 3rd order:	[N] are now the [*1] popular [N].
Template 3rd order (variation):	[N] is now the [*1] popular [N].
*1. {most, least}	

Category:	Circumstance
Reference:	R17
Original Sentence:	When [the Pew Internet Personal Networks and Community] survey was conducted (July 9-August 10, 2008), [77%] of the [U.S. adult population] used [the internet].
Template 1st order:	When [M] survey was conducted, [B] per cent of the [C] [H] used O.
Template 2nd order:	When [M] survey was conducted, [E] ,[B] per cent of the [H] used O.

Category:	Circumstance
Reference:	R18
Original Sentence:	In [2004], just [18%] of [12] year olds had a cell phone of their own.
Template 1st order:	In [E], just [B] per cent of [A] year olds had [N].
Template 2nd order:	In [E], just [B] per cent of [H] had [N].

Category:	Circumstance
Reference:	R18
Original Sentence:	As of [December 2009], [74%] of [adults] use the internet.
Template 1st order:	As of [E], [B] per cent of [H] use the internet.
Template 2nd order:	As of [E], [B] per cent of [H] use [O].
Template 2nd order (variation):	As of [E], [B] per cent of [H] use [N].

Category:	Circumstance
Reference:	R15
Original Sentence:	Average [Twitter] usage frequency is [23] times [a week].
Template 1st order:	Average [O] usage frequency is [A] times [D].
Template 2nd order:	Average [N] usage frequency is [A] times [D].

Category:	Circumstance
Reference:	R15
Original Sentence:	[46%] of [14]- [19] years olds access [social networking sites] [every day].
Template 1st order:	[B] per cent of [A]- [A] years olds access social networking sites [D].
Template 2nd order:	[A] years olds access [N] [D].

Category:	Circumstance
Reference:	R18
Original Sentence:	[47%] of [online] [adults] use [social networking sites], [up] from [37%] in [November 2008].
Template 1st order:	[B] per cent of online social networking sites use [O], [*1] from [B] per cent in [E].
Template 2nd order:	[B] per cent of online [H] use [O], [*1] from [B] per cent in [E].
Template 2nd order (variation):	[B] per cent of online [H] use [N], [*1] from [B] per cent in [E].
*1. {up, down}	

Category:	Circumstance
Reference:	R16
Original Sentence:	[8%] of online [adults] who do not currently use [Facebook] are interested in becoming [Facebook] users in [the future].
Template 1st order:	[B] per cent of online [H] who do not currently use [O] are interested in becoming [O] users in [D].
Template 2nd order:	[H] are interested in [N].

Category:	Circumstance
Reference:	R17
Original Sentence:	[Internet users] are [26%] [less] likely [to rely] on their neighbors for help with small services.
Template 1st order:	Internet users are [B] per cent [*1] likely to rely on [H].
Template 1st order (variation):	[H] are [B] per cent [*1] likely to rely on [H].
Template 2nd order:	[H] are [B] per cent [*1] likely [L].
*1. {more, less}	

Category:	Circumstance
Reference:	R20
Original Sentence:	At that time just [8%] of [internet] [users] or [5%] of all [adults] said they used them.
Template 1st order:	At that time just [B] per cent of internet users or [B] per cent of all [H] said they use social networking sites.
Template 2nd order:	At that time just [B] per cent of [H] said they use [N].

5.1.4. De justificación

Category:	Justify
Reference:	R4
Original Sentence:	[The 'National Perspectives'] report from [the EU Kids Online project based at the London School of Economics and Political Science (LSE)] reveals that about [91%] of [UK children] go [online] [at school] when compared to European average of [63%].
Template 1st order:	[M] reveals that about [B] per cent of [H] go [*1] at [C].
*1. {online, offline}	

Category:	Justify
Reference:	R4
Original Sentence:	The report revealed that on an average, [teen agers] between nine and 16 years old spend [102 minutes] on [the internet] when in European countries, the average time spent on internet is about 88 minutes.
Template 1st order:	The report revealed that on an average, [H] spend [F] on [O].
Template 2nd order:	[M] revealed that on an average, [H] spend [F] on [N].

Category:	Justify
Reference:	R6
Original Sentence:	A new national survey by [the Pew Research Center's Internet & American Life Project] has found that [75%] of all [American adults] are active in [some kind of voluntary group or organization] and [internet] users are more likely than others to be active.
Template 1st order:	A new survey by [M] has found that [B] per cent of all [H] are active in some kind of group.
Template 2nd order:	A new survey by [M] has found that [B] per cent of all [H] are active in [O].

5.1.5. De evaluación

Category:	Evaluation
Reference:	R14
Original Sentence:	Only [one] in [five] [teenagers] and [one] in [twenty] [adults] said that [people] were ["mostly unkind"] on [social networking sites] such as [Facebook] and [Twitter].
Template 1st order:	Only [A] in [A] [H] said that [H] were [I] on social networking sites.

5.1.6. De resultado voluntario

Category:	Volitional Result
Reference:	R3
Original Sentence:	[25%] of [social media] [teens] have had an experience on [a social network site] that resulted in [a face-to-face argument] or [confrontation] with [someone].
Template 1st order:	[B] per cent of [H] have had an experience on [N] that resulted in [N].
Template 2nd order:	[B] per cent of [H] have had an experience on [N].

5.1.7. De elaboración

Category:	Elaboration
Reference:	R15
Original Sentence:	Among [internet users], some [62%] use [social networking sites] such as [Facebook], [Twitter], [MySpace] or [Linkedin].
Template 1st order:	Among internet users, some [B] per cent use [social networking sites] such as [O] or [O].
Template 2nd order:	[B] per cent of [H] use [N].

Category:	Elaboration
Reference:	R18
Original Sentence:	Fully [80%] of [teens] between the ages of [12] and [17] have a game console like a Wii, an Xbox or a PlayStation.
Template 1st order:	Fully [B] per cent of [H] between the ages of [A] and [A] have [N].
Template 2nd order:	[H] between the ages of [A] and [A] have [N].

Category:	Elaboration
Reference:	R15
Original Sentence:	[Facebook] dominates as the [most] used [social networking site], being used by [97%] of [social networking participants].
Template 1st order:	[O] dominates as the [*1] used social networking site, being used by [B] per cent of social networking participants.
Template 2nd order:	[N] dominates as the [*1] used [N].
*1. {most, least}	

5.2. Análisis de textos aplicando la teoría RST

5.2.1. Texto 1

Social Networking

As of December 2012:

- 15% of online adults say they use Pinterest.
- 13% of online adults say they use Instagram.
- 6% of online adults say they use Tumblr.
- 67% of online adults say they use Facebook.
- 16% of online adults say they use Twitter.
- 20% of online adults say they use LinkedIn as of August 2012.

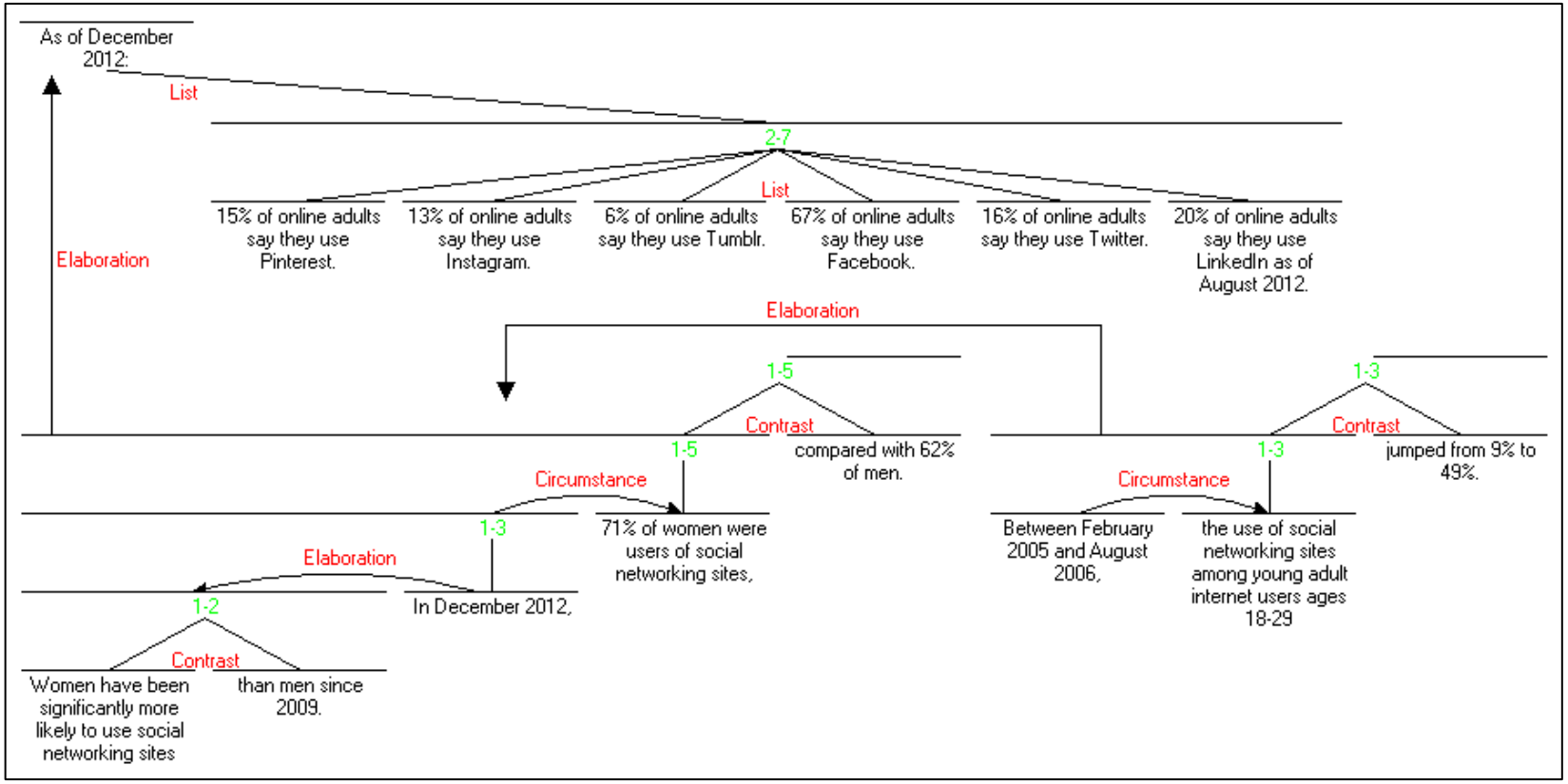
Women have been significantly more likely to use social networking sites than men since 2009. In December 2012, 71% of women were users of social networking sites, compared with 62% of men.

Between February 2005 and August 2006, the use of social networking sites among young adult internet users ages 18-29 jumped from 9% to 49%.

Resultados obtenidos:

Relación	Nº	Media
Circumstance	2	13.3%
Contrast	3	20%
Elaboration	3	20%
List	7	46.7%

Análisis RST:



5.2.2. Texto 2

28% of nine and ten-year-olds in UK use social networking websites: study

UK children are more expected to go online through a mobile or handheld device

About 28% of UK 9-10 year olds and 59% of 11-12 year olds operate a social networking profile, though social networking service (SNS) including Facebook have a minimum age of 13 years, according to a new study.

The 'National Perspectives' report from the EU Kids Online project based at the London School of Economics and Political Science (LSE) reveals that about 91% of UK children go online at school when compared to European average of 63%.

The report revealed that on an average, teen agers between nine and 16 years old spend 102 minutes on the internet when in European countries, the average time spent on internet is about 88 minutes.

EU Kids Online project at LSE senior researcher Dr Leslie Haddon said the report includes findings for 33 European countries, allowing direct comparisons in the experiences of children as they go online in different countries.

"These national differences mean that there is no one-size-fits-all-solution for children's internet safety," Haddon said.

The report also revealed that UK children are more expected to go online through a mobile or handheld device, placing them as precursors of new risks connected with personal internet access and making defensive supervision by their parents even more difficult.

LSE Professor Sonia Livingstone said EU Kids Online has categorized the UK as a 'high use, some risk' country, an improvement on previous findings of 'high use, high risk'.

"It seems that the considerable multi-stakeholder efforts are bearing fruit," Livingstone said.

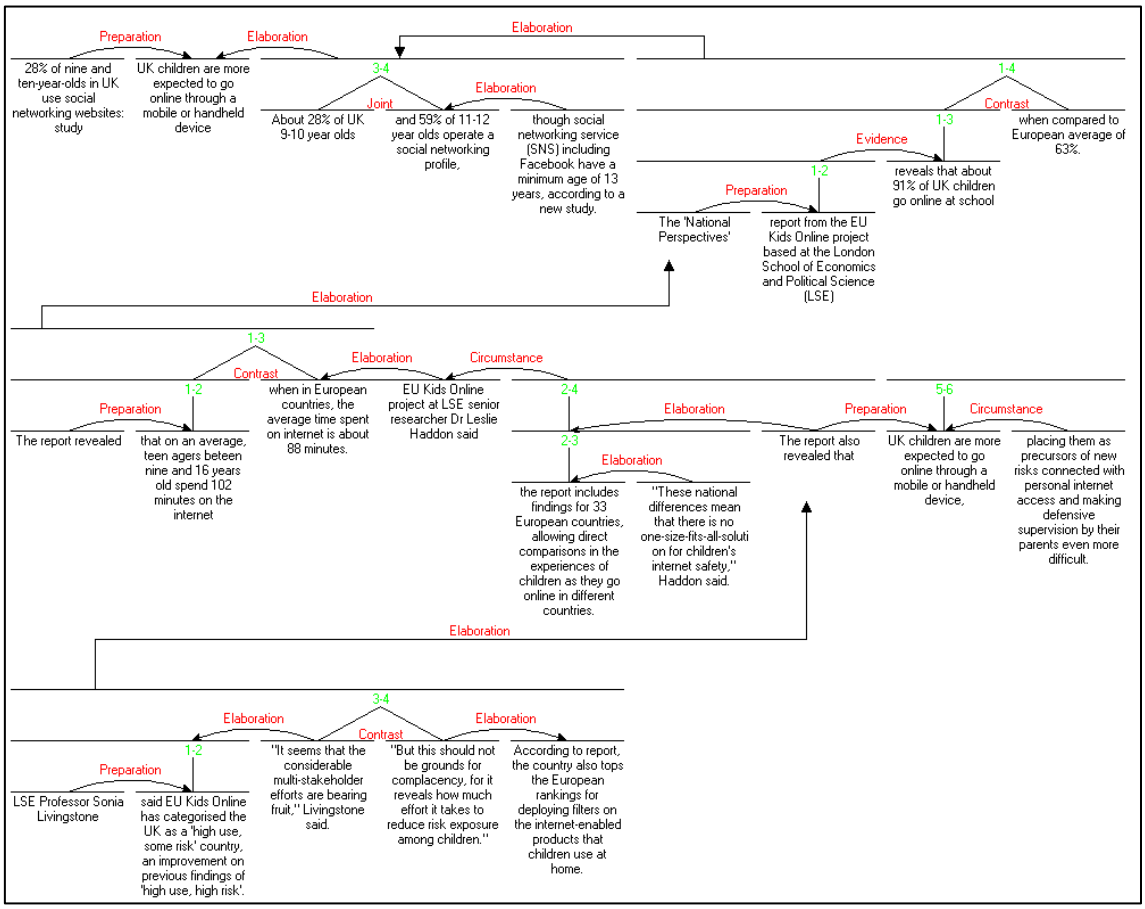
"But this should not be grounds for complacency, for it reveals how much effort it takes to reduce risk exposure among children."

According to report, the country also tops the European rankings for deploying filters on the internet-enabled products that children use at home.

Resultados obtenidos:

Relación	Nº	Media
Circumstance	2	9.09%
Contrast	3	13.63%
Elaboration	10	45.45%
Evidence	1	4.54%
Joint	1	4.54%
Preparation	5	22.72%

Análisis RST:



5.2.3. Texto 3

65% of online adults use social networking sites

Fully 65% of adult internet users now say they use a social networking site like MySpace, Facebook or LinkedIn, up from 61% one year ago. This marks the first time in Pew Internet surveys that 50% of all adults use social networking sites.

These figures come from a new national survey by the Pew Research Center's Internet & American Life Project and mark a dramatic increase from the first time the Project surveyed about social networking sites in February of 2005. At that time just 8% of internet users or 5% of all adults said they used them.

Among internet users, social networking sites are most popular with women and young adults, but most of the growth over the past year came from adults over age 30. Looking at overall usage, wired seniors grew their ranks the most over the past year; 33% of those ages 65 and older now use the sites, compared with 26% one year ago.

As of May 2011:

- 83% of internet users ages 18-29 use SNS, compared with
- 70% of 30-49 year-olds
- 51% of 50-64 year-olds, and
- 33% of those ages 65 and older

Looking at usage on a typical day, 43% of online adults use social networking, up from 38% a year ago. Out of all the "daily" online activities that we ask about, only email (which 61% of internet users access on a typical day) and search engines (which 59% use on a typical day) are used more frequently than social networking tools.

The frequency of social networking site usage among young adult internet users was stable over the last year – 61% of online Americans in that age cohort now use SNS on a typical day, compared with 60% one year ago. However, among the Boomer-aged segment of internet users ages 50-64, SNS usage on a typical day grew a significant 60% (from 20% to 32%).

"The graying of social networking sites continues, but the oldest users are still far less likely to be making regular use of these tools," said Mary Madden, a Senior Research Specialist with the Project and co-author of the report. "While seniors are testing the waters, many Baby Boomers are beginning to make a trip to the social media pool part of their daily routine," said Madden.

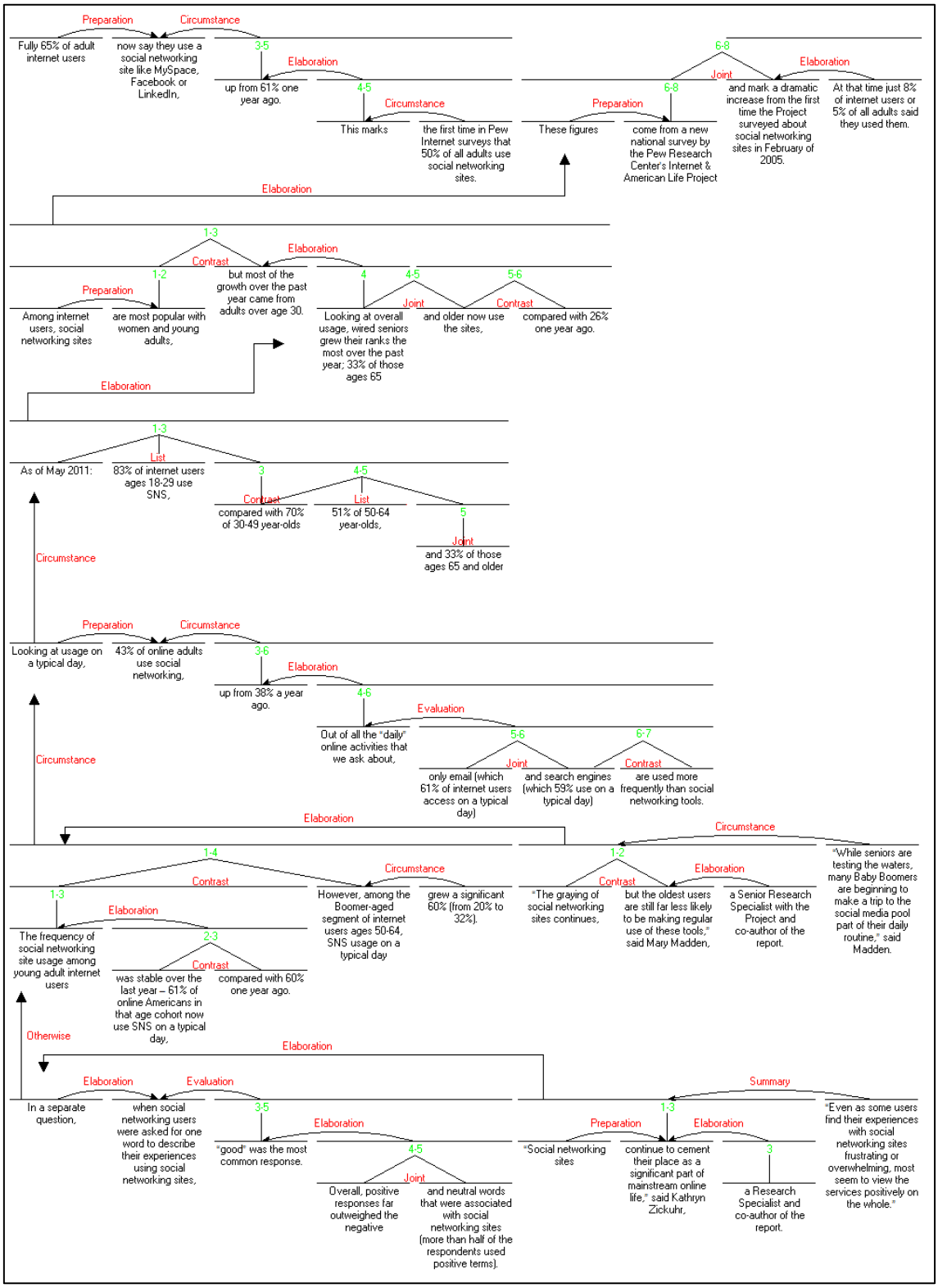
In a separate question, when social networking users were asked for one word to describe their experiences using social networking sites, "good" was the most common response. Overall, positive responses far outweighed the negative and neutral words that were associated with social networking sites (more than half of the respondents used positive terms). Users repeatedly described their experiences as "fun," "great," "interesting" and "convenient." Less common were superlatives such as "astounding," "necessity," and "empowering."

"Social networking sites continue to cement their place as a significant part of mainstream online life," said Kathryn Zickuhr, a Research Specialist and co-author of the report. "Even as some users find their experiences with social networking sites frustrating or overwhelming, most seem to view the services positively on the whole."

Resultados obtenidos:

Relación	Nº	Media
Circumstance	5	11.11%
Contrast	9	20%
Elaboration	11	24.44%
Evaluation	2	4.44%
Joint	5	11.11%
List	6	13.33%
Preparation	5	11.11%
Summary	1	2.22%
Otherwise	1	2.22%

Análisis RST:



5.2.4. Texto 4

Do we leave our manners outside of social media?

Perhaps it's because online comments can become a catalyst for a mob mentality and is often set in stone digitally -- rather than being lost and forgotten if it's verbal -- every small action can seem worse online.

Either that or such transparent and blatant abuse we can find on social networks means we remember it more often -- with evidence suggesting many will act in a fashion that wouldn't be tolerated in the physical world under the cloak of 'anonymity'.

According to a recent study conducted by Pew and interpreted by Salesforce Rypple in the infographic below, only one in five teenagers and one in twenty adults said that people were "mostly unkind" on social networking sites such as Facebook and Twitter.

69 and 85 percent stated that people were "mostly kind" respectively, and some even use the networks in order to promote their self-esteem. Over half of children and adults attribute social media as a means to better their self-confidence,

However, as we may expect, networks that teenagers tend to frequent are often a base for negative behavior. When teenagers become involved in these situations online, the consequences physically were said to be:

- 25 percent later got into trouble at work or school;
- 22 percent became involved in a physical fight;
- 13 percent said that social networking caused issues with their family;
- 8 percent had a face-to-face argument;

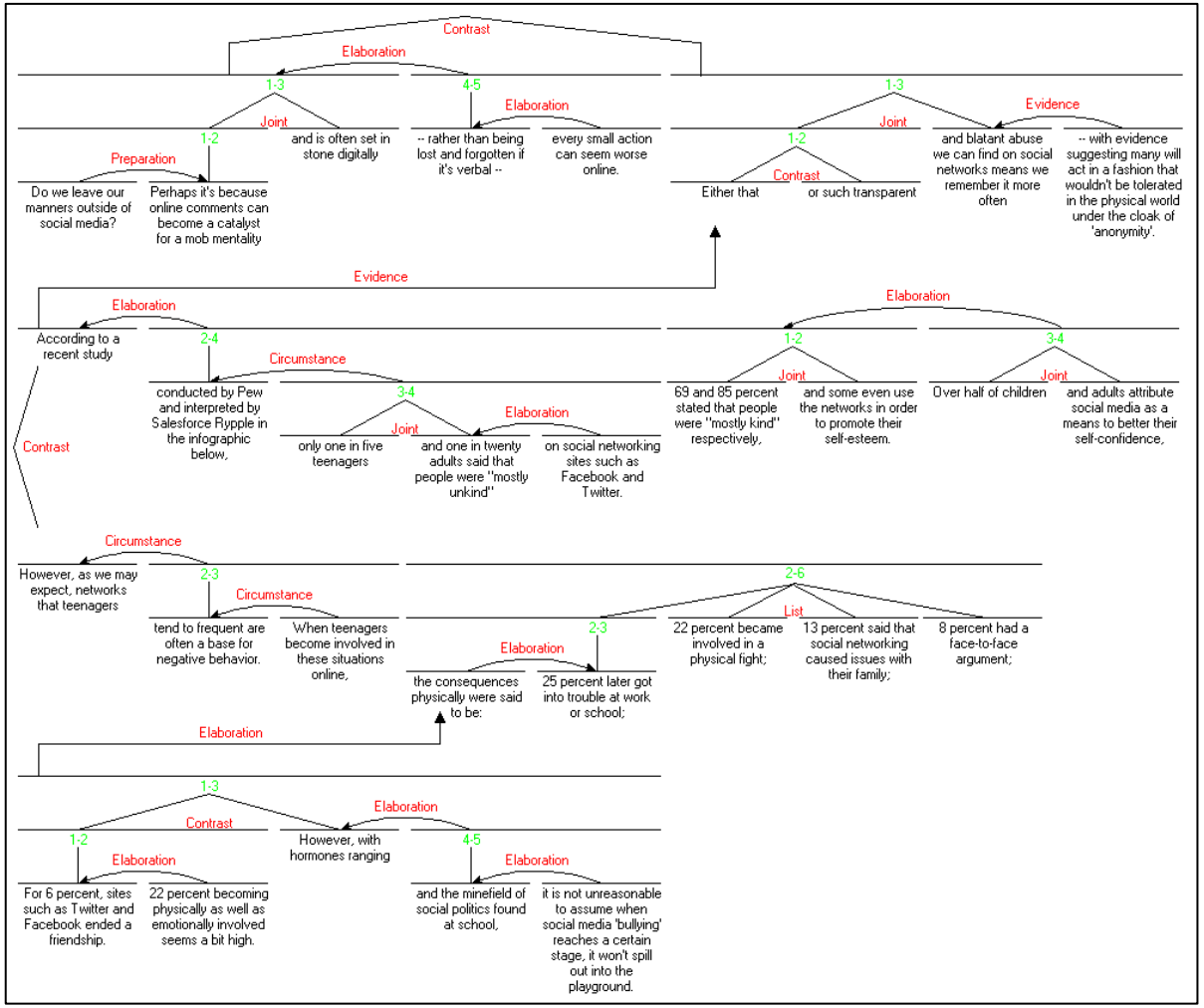
For 6 percent, sites such as Twitter and Facebook ended a friendship.

22 percent becoming physically as well as emotionally involved seems a bit high. However, with hormones ranging and the minefield of social politics found at school, it is not unreasonable to assume when social media 'bullying' reaches a certain stage, it won't spill out into the playground.

Resultados obtenidos:

Relación	Nº	Media
Circumstance	3	10.34%
Contrast	4	13.79%
Elaboration	10	34.48%
Evidence	2	6.89%
Joint	5	17.24%
List	4	13.79%
Preparation	1	3.44%

Análisis RST:



5.2.5. Texto 5

Questions have been raised about the social impact of widespread use of social networking sites (SNS) like Facebook, LinkedIn, MySpace, and Twitter. Do these technologies isolate people and truncate their relationships? Or are there benefits associated with being connected to others in this way? The Pew Research Center's Internet & American Life Project decided to examine SNS in a survey that explored people's overall social networks and how use of these technologies is related to trust, tolerance, social support, and community and political engagement.

The findings presented here paint a rich and complex picture of the role that digital technology plays in people's social worlds. Wherever possible, we seek to disentangle whether people's varying social behaviors and attitudes are related to the different ways they use social networking sites, or to other relevant demographic characteristics, such as age, gender and social class.

The number of those using social networking sites has nearly doubled since 2008 and the population of SNS users has gotten older.

In this Pew Internet sample, 79% of American adults said they used the internet and nearly half of adults (47%), or 59% of internet users, say they use at least one of SNS. This is close to double the 26% of adults (34% of internet users) who used a SNS in 2008. Among other things, this means the average age of adult-SNS users has shifted from 33 in 2008 to 38 in 2010. Over half of all adult SNS users are now over the age of 35. Some 56% of SNS users now are female.

Facebook dominates the SNS space in this survey: 92% of SNS users are on Facebook; 29% use MySpace, 18% used LinkedIn and 13% use Twitter.

There is considerable variance in the way people use various social networking sites: 52% of Facebook users and 33% of Twitter users engage with the platform daily, while only 7% of MySpace and 6% of LinkedIn users do the same.

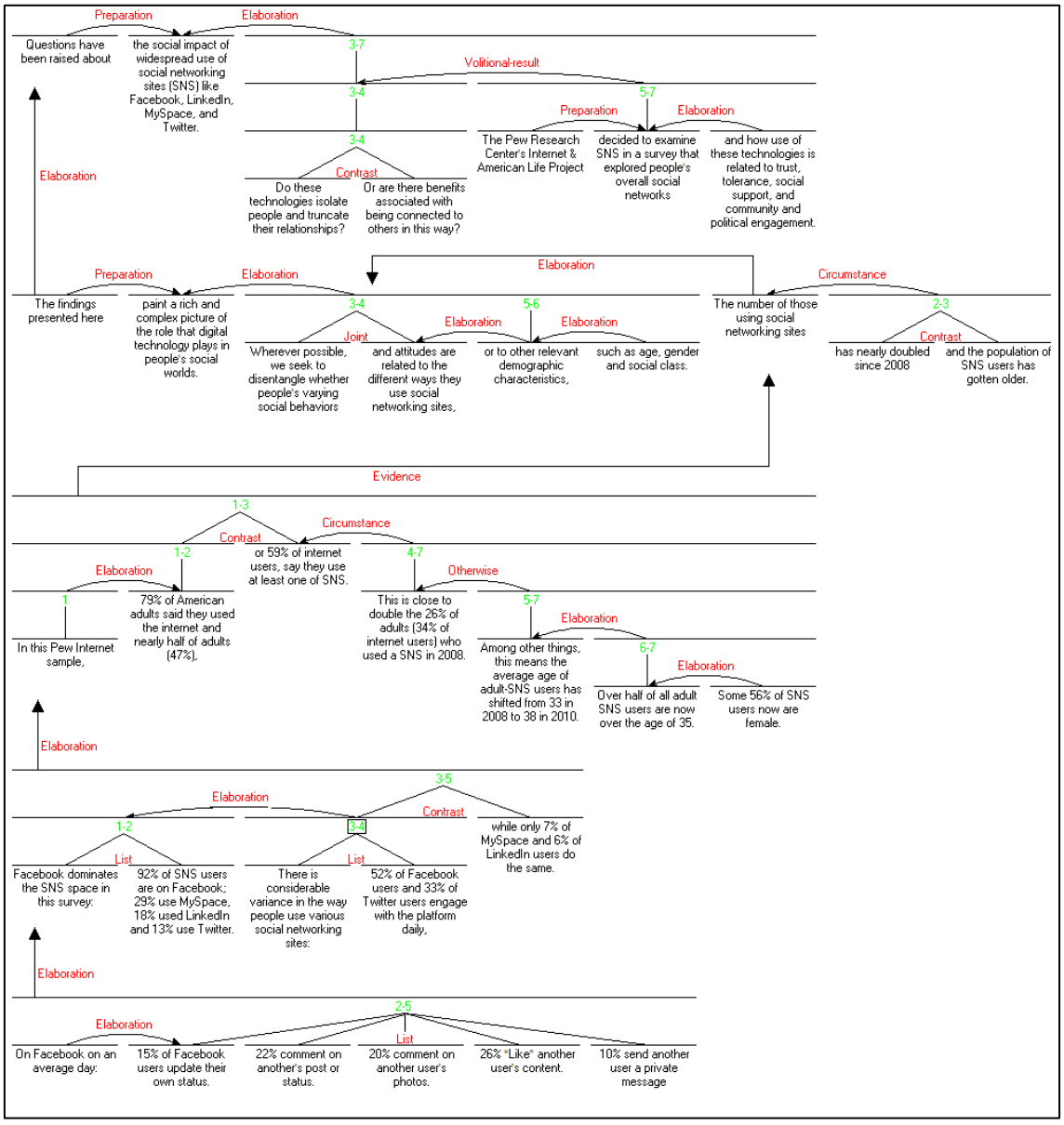
On Facebook on an average day:

- 15% of Facebook users update their own status.
- 22% comment on another's post or status.
- 20% comment on another user's photos.
- 26% "Like" another user's content.
- 10% send another user a private message

Resultados obtenidos:

Relación	Nº	Media
Circumstance	2	6.06%
Contrast	4	12.12%
Elaboration	14	42.42%
Evidence	1	3.03%
Joint	1	3.03%
List	6	18.18%
Preparation	3	9.09%
Otherwise	1	3.03%
Volitional Result	1	3.03%

Análisis RST:



5.2.6. Texto 6

The report is the first national survey of how the use of social networking sites (SNS) by adults is related to people's overall social networks. The findings suggests that there is little validity to concerns that people who use SNS experience smaller social networks, less closeness, or are exposed to less diversity. We did find that people who are already likely to have large overall social networks – those with more years of education – gravitate to specific SNS platforms, such as LinkedIn and Twitter. The size of their overall networks is no larger (or smaller) than what we would expect given their existing characteristics and propensities.

However, total network size may not be as important as other factors – such as intimacy. Americans have more close social ties than they did two years ago. And they are less socially isolated. We found that the frequent use of Facebook is associated with having more overall close ties.

In addition, we found that only a small fraction of Facebook friends are people whom users have never met or met only once.

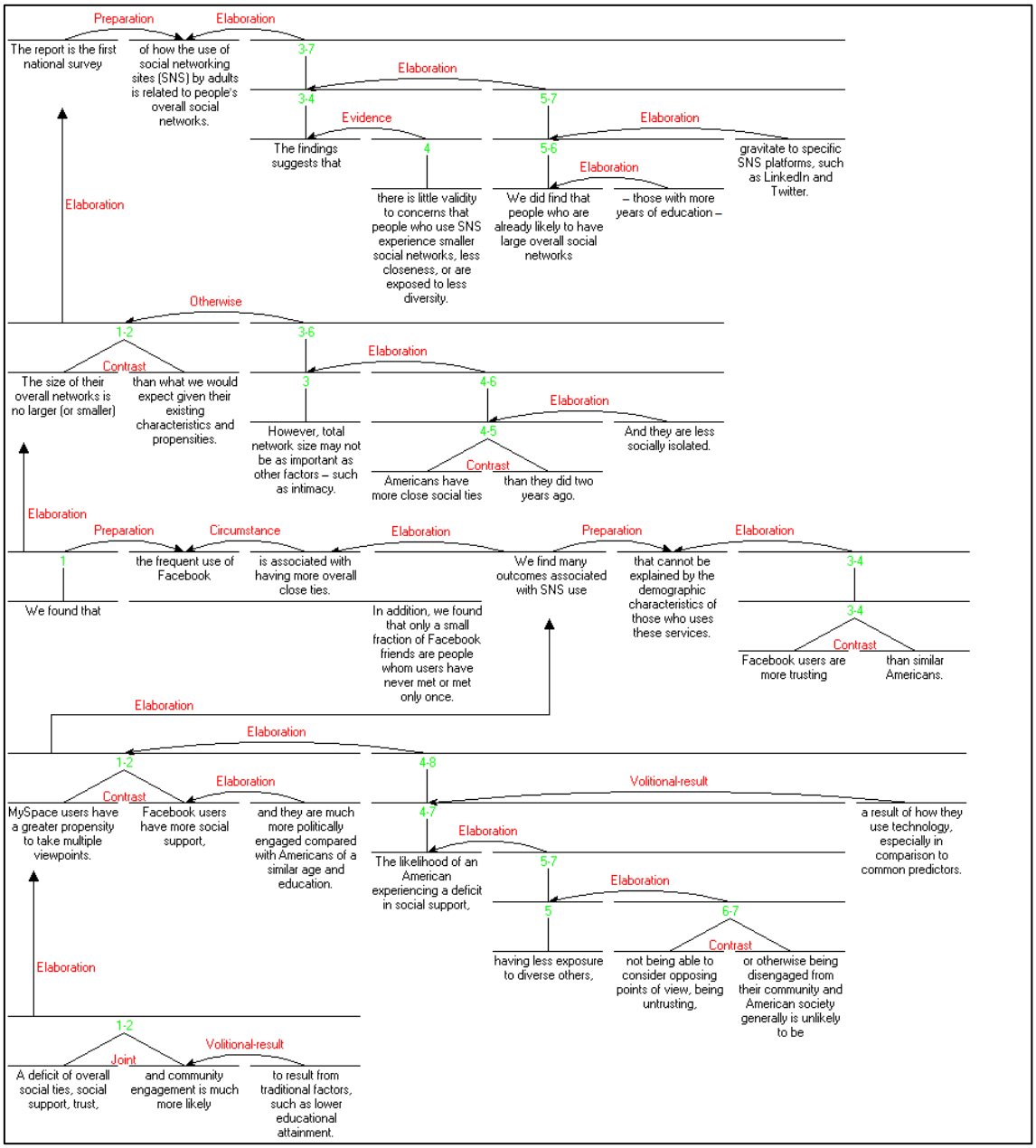
We find many outcomes associated with SNS use that cannot be explained by the demographic characteristics of those who uses these services. Facebook users are more trusting than similar Americans. MySpace users have a greater propensity to take multiple viewpoints. Facebook users have more social support, and they are much more politically engaged compared with Americans of a similar age and education.

The likelihood of an American experiencing a deficit in social support, having less exposure to diverse others, not being able to consider opposing points of view, being untrusting, or otherwise being disengaged from their community and American society generally is unlikely to be a result of how they use technology, especially in comparison to common predictors. A deficit of overall social ties, social support, trust, and community engagement is much more likely to result from traditional factors, such as lower educational attainment.

Resultados obtenidos:

Relación	Nº	Media
Circumstance	1	3.33%
Contrast	5	16.66%
Elaboration	16	53.33%
Evidence	1	3.33%
Joint	1	3.33%
Preparation	3	10%
Otherwise	1	3.33%
Volitional Result	2	6.66%

Análisis RST:



5.3. Referencias de textos

R1 http://www.computerworld.com.au/article/438052/enterprise_social_networking_market_heats_up_ovum/

R2 http://pewinternet.org/Commentary/2012/March/Pew-Internet-Social-Networking-full-detail.aspx

R3 http://www.pewinternet.org/~/media//Files/Reports/2011/PIP_Teens_Kindness Cruelty_SNS_Report_Nov_2011_FINAL_110711.pdf

R4 http://media.cbronline.com/news/28-of-nine-and-ten-year-olds-in-uk-use-social-networking-websites-study-171012

R5 http://www.pewinternet.org/Reports/2011/Technology-and-social-networks/Summary.aspx

R6 http://www.pewinternet.org/Reports/2011/The-Social-Side-of-the-Internet/Summary.aspx

R7 http://www.pewinternet.org/Reports/2010/Neighbors-Online/Part-1.aspx

R8 http://www.pewinternet.org/Reports/2009/15--The-Internet-and-Civic-Engagement/1--Summary-of-Findings.aspx

R9 http://www.pewinternet.org/Reports/2006/The-Strength-of-Internet-Ties/01-Summary-of-Findings.aspx

R10 http://computer.howstuffworks.com/internet/social-networking/information/social-networks-honesty.htm

R11 http://www.usatoday.com/story/news/nation/2013/04/18/social-media-tweens-fame/2091199/

R12 http://www.digitaltrends.com/social-media/cheaters-use-social-networks/

R13 http://techcrunch.com/2011/09/14/over-1-billion-people-use-social-networks-today-and-other-stats/

R14 http://www.zdnet.com/blog/igeneration/do-we-leave-our-manners-outside-of-social-media/16412

R15 http://about.sensis.com.au/IgnitionSuite/uploads/docs/FinalYellow_SocialMedia Report_digital_screen.pdf

R16 http://www.pewinternet.org/~/media//Files/Reports/2013/PIP_Coming_and_going_on_facebook.pdf

R17 http://www.pewinternet.org/~/media//Files/Reports/2009/PIP_Tech_and_Social_Isolation.pdf

R18 http://www.pewinternet.org/~/media//Files/Reports/2010/PIP_Social_Media_and_Young_Adults_Report_Final_with_toplines.pdf

R19 http://www.pewinternet.org/~/media//Files/Reports/2011/PIP%20-%20Social%20networking%20sites%20and%20our%20lives.pdf

R20 http://www.pewinternet.org/Press-Releases/2011/65-of-online-adults-use-social-networking-sites.aspx

BIBLIOGRAFÍA

BIBLIOGRAFÍA

[Bradshaw, 2013] P. Bradshaw. (2013, April, 11). *What is Data Journalism? (1st ed.)* [Online]. Available: http://datajournalismhandbook.org/1.0/en/introduction_0.html

[Busacca, 1998] E. Busacca. (1998, Feb, 02). *Eliza (1st ed.)* [Online]. Available: http://www.letralia.com/40/ar01-040.htm

[Carbonell, 1992] J. Carbonell. (1992). *El procesamiento del lenguaje natural, tecnología en transición (1st ed.)* [Online]. Available: http://cvc.cervantes.es/obref/congresos/sevilla/tecnologias/ponenc_carbonell.htm

[EPSRC, 2005] EPSRC. (2005, Nov).*EPSRC Funded project for generating Summaries of Time Series Data (last ed.)* [Online]. Available: http://inf.abdn.ac.uk/research/sumtime/

[Harnad, 2005] S. Harnad. (2005 Mar). *The Implementation of the Berlin Declaration on Open Access (last ed.)* [Online]. Available: http://www.dlib.org/dlib/march05/harnad/03harnad.html

[Mann, 2006] W. C. Mann, M. Taboada. (10, July, 2006). *Applications of Rhetorical Structure Theory (2nd ed.)* [Online]. Available: http://www.sfu.ca/~mtaboada/docs/Taboada_Mann_RST_Part2.pdf

[Mann, 2006] W. C. Mann, S. A. Thompson. (23, Jan, 2006). *Rheteorical Structure Theory: Toward a functional theory of text organization (1st ed.)* [Online]. Available: http://www.cis.upenn.edu/~nenkova/Courses/cis700-2/rst.pdf

[Mann, 2012] W. C. Mann, M. Taboada. (2012). *Introducción a la Teoría de la Estructura Retórica (last ed.)* [Online]. Available: http://www.sfu.ca/rst/08spanish/introduccion.html

[Moncur, 2008] W. Moncur (2008). *From Data to Text in the Neonatal Intensive Care Unit: Using NLG Technology for Decision Support and Information Management (1st ed)* [Online]. Available: http://www.academia.edu/461807/From_Data_to_Text_in_the_Neonatal_Intensive_Care_Unit_Using_NLG_Technology_for_Decision_Support_and_Information_Management

[Nevin, 2002] B.E. Nevin, S.M. Johnson "Text Generations within sublanguages" in *The Legacy of Zellig Harris*, vol. 2, Ed:John Benjamins Publishing Co., 2002, pp 241-242.

[O'Donnell, 2006] M. O'Donnell. (2006). *RSTTool -- an RST Markup Tool (3.0 ed)* [Online]. Available: http://www.wagsoft.com/RSTTool/index.html

[OKF, 2012] OFK. (2012). *What is Open Data?(1.1 ed.)* [Online]. Available: http://okfn.org/opendata/

[OPL, 1998] OPL. (1998, Jul, 14). *Open Content (last ed.)* [Online]. Available: http://opencontent.org/opl.shtml

[OSI, 1999] OSI. (1999). *The Open Source Definition (Annotated) (1.9 ed)* [Online]. Available: http://opensource.org/osd-annotated

[Reiter, 2007] E. Reiter. (2007). *An architecture for data-to-text systems (1.3 ed)* [Online]. Available: http://dl.acm.org/citation.cfm?id=1610180

[Saenz, 2010] A.Saenz. (2010, Jan, 13). *Cleverbot chat engine is learning from the internet to talk like a human (1st ed.)* [Online]. Available: http://singularityhub.com/2010/01/13/cleverbot-chat-engine-is-learning-from-the-internet-to-talk-like-a-human/

[Suehle, 2011] R. Suehle. (2011, Feb). *First open hardware definition has been released (last ed.)* [Online]. Available: http://opensource.com/life/11/2/first-open-hardware-definition-has-been-released

[Wallace, 2013] R.S. Wallace. (2013, April, 9). *Introducing ALICE 2.0 (last ed.)* [Online]. Available: http://www.alicebot.org/

Este documento esta firmado por

	**Firmante**	CN=tfgm.fi.upm.es, OU=CCFI, O=Facultad de Informatica - UPM, C=ES
	Fecha/Hora	Fri Feb 14 19:23:07 CET 2014
	Emisor del Certificado	EMAILADDRESS=camanager@fi.upm.es, CN=CA Facultad de Informatica, O=Facultad de Informatica - UPM, C=ES
	Numero de Serie	630
	Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)